Dpto. de Teoría de la Señal y Comunicaciones
Universidad Carlos iii de Madrid

TESIS DOCTORAL

EXTRACCIÓN DE CARACTERÍSTICAS MEDIANTE CRITERIOS BASADOS EN TEORÍA DE LA INFORMACIÓN

Autor: Jose Miguel Leiva Murillo
Director: Dr. Antonio Artés Rodríguez

Leganés, 2007

Tesis Doctoral:
EXTRACCIÓN DE CARACTERÍSTICAS MEDIANTE CRITERIOS BASADOS EN TEORÍA DE LA INFORMACIÓN.

Autor:
JOSÉ MIGUEL LEIVA MURILLO

Director:
DR. ANTONIO ARTÉS RODRÍGUEZ

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores

Presidente:

Vocales:

Secretario:

acuerda otorgarle la calificación de

Leganés, a

RESUMEN

El objetivo de esta tesis es desarrollar nuevas técnicas de extracción lineal de características para clasificación, mediante el uso de criterios basados en teoría de la información. Dado que esta teoría establece un marco que permite medir el grado de dependencia entre variables aleatorias o señales, nos proporciona una buena medida del nivel de relevancia de las características extraídas respecto a la clase a la que pertenece cada muestra. Se exploran dos aproximaciones distintas. La primera de ellas está basada en funciones de contraste que aproximan la información mutua entre las proyecciones obtenidas individualmente y las clases. En la segunda, se plantea como criterio a optimizar el test de hipótesis que sirve como regla de clasificación.

La aproximación basada en funciones de contraste hace uso de una estimación de la información mutua similar a otra que ha sido usada con éxito en análisis de componentes independientes. La estimación utiliza momentos sobre funciones no polinomiales aplicadas a los datos. De esta forma, no es necesario el modelo de la función densidad de probabilidad (FDP) de los datos. Al ser derivables estas funciones, es posible su optimización mediante ascenso por gradiente.

El método basado en la maximización del test de hipótesis hace uso de un modelo de Parzen para estimar la FDP de los datos. Se propone un algoritmo para optimizar dichos modelos de acuerdo con el criterio de máxima verosimilitud combinado con un procedimiento de validación cruzada. Su aplicación al problema de extracción lineal de características consiste en la maximización del cociente de verosimilitudes medido sobre los datos de entrenamiento. También se describen procedimientos para la aplicación de estos modelos optimizados a otros problemas de aprendizaje como análisis de componentes independientes y clasificación de Parzen.

Junto a los métodos propuestos, se presentan los resultados de la aplicación de los mismos a problemas reales, registrados en bases de datos públicas. La tesis finaliza con las conclusiones más relevantes que pueden extraerse del trabajo presentado, así como la propuesta de líneas de trabajo que constituyen la continuación natural de las ya expuestas.

ABSTRACT

The objective of this PhD Thesis is to develop new linear feature extraction methods for classification, by the use of criteria from Information Theory. This theory allows us to measure the statistical dependence among random variables or signals; in a classification context, the measure of interest is the relevance of the feature extracted with respect to the labels. Two different methodologies are explored: the first one is based on contrast functions that approximate the mutual information between the set of projections obtained and the classes. The second approach proposes a hypothesis test scheme as a rule for optimal classification performance.

The approach based on contrast functions makes use of a mutual information estimation related to another one that has been applied to Independent Component Analysis (ICA) with success. The estimation is based on non polynomial moments obtained from data. This way, the probability density function (PDF) need not to be estimated. As the non polynomial functions are derivable, an optimization procedure by gradient ascent is applied.

The approach based on the hypothesis test maximization does make use of the PDF from the data, by means of Parzen non-parametric models, also known as kernel density estimators. To optimize these models, an algorithm is proposed for determining the optimal window width by a maximum likelihood criterion, combined with a cross validation procedure. The application of these models to the feature extraction problem is based on the maximization of the hypothesis test among classes estimated from the training set, using the features extracted. In addition to feature extraction, the Thesis proposes methods for the application of these optimized PDF modelos to other machine learning problems as ICA and Parzen classification.

Simulation results are provided, in which the methods proposed are applied to real problems, recorded on public datasets. The Thesis finishes with some conclusions about the advantages and weaknesses of each procedure. Some suggestions are also made about frameworks that can be considered as the natural continuation of the work presented.

AGRADECIMIENTOS

Quiero dar las gracias en primer lugar al Profesor Dr. Antonio Artés por haberme dado la posibilidad de formarme en el Departamento de Teoría de la Señal, un entorno de gran calidad intelectual y humana. Le agradezco su acierto en la temática elegida para la tesis, así como su paciente supervisión.

Este periodo de formación arrancó en el laboratorio 4.2.A03, que en mis primeros años de doctorado albergaba una densidad de talento difícil de igualar, y una calidad humana imposible de superar. Al profesor Sancho Salcedo le agradezco los estímulos que nos hacía llegar al resto de doctorandos y su capacidad para encontrar aplicaciones cotidianas a las técnicas que manejamos habitualmente. A Ricardo Torres le doy las gracias por haber compartido con nosotros su brillantez, su bondad y (por qué no decirlo) su acidez. José Emilio Vila duró poco en ese laboratorio, pero ahora que ha vuelto ha sido un gran apoyo en esta interminable recta final. Con Mario de Prado he compartido mucho durante estos años; en lo profesional, y sobre todo en lo personal, le estoy en deuda por su disposición eterna a ayudar. A Ricardo Santiago le doy un millón de gracias por su ayuda con cualquier cosa que exija virtuosismo (Emacs, Latex, Matlab, Linux, C, etc, etc), sus sugerencias e ideas en lo científico, y su gran generosidad en cosas relacionadas o no con la tesis.

Le estoy muy agradecido a Harold Molina por su ayuda prestada cuando los experimentos fallaban (a veces no era culpa mía), y a Ascensión Gallardo por su capacidad de trabajo en los artículos en que hemos colaborado. También agradezco a Matilde Sánchez, Fernando Pérez, Marcelino Lázaro, Emilio Parrado y Jerónimo Arenas su compañerismo y tantos buenos ratos.

A mis padres, Juan y Josefina, a mi abuela Joaquina y a mis hermanos, Quini y M^{a} del Mar les agradezco su permanente interés en mi trabajo y su apoyo a todo lo que haga, aunque sea seguir estudiando hasta los 30. A ellos, y a todos los demás que no tienen claro para qué sirve una tesis, les debo una respuesta.

Le dedico la tesis a Mari Carmen, que ha sido la que más me ha animado, y la que más de cerca ha visto mi esfuerzo.

Índice general

Capítulo 1

Introducción

Las actuales técnicas de minería de datos, aprendizaje estadístico o reconocimiento de patrones plantean aplicaciones cada vez más ambiciosas de extracción de la información relevante contenida en los datos. Además, el volumen de éstos es cada vez mayor, lo que plantea diversos retos en cuanto al coste computacional de su procesado. Por esta razón, los métodos de selección y de extracción de características han adoptado un papel protagonista en el ámbito del aprendizaje máquina.

Un procedimiento de aprendizaje basado en ejemplos puede ser descrito a partir de tres procesos que tienen lugar secuencialmente según se muestra en la Figura 1.1 [Bishop, 1995] [Duda, 2000]. Cada bloque desempeña las siguientes funciones:

1. El **preprocesado** lleva a cabo una serie de operaciones que toman como entrada las magnitudes medidas por los sensores que correspondan, y presentan a su salida unos valores numéricos susceptibles de ser procesados e interpretados.

2. El objetivo de la **extracción de características** es eliminar la redundancia presente en los datos y proporcionar al siguiente módulo un conjunto de valores lo más representativo y útil como sea posible para la tarea posterior. Con frecuencia, este bloque recibe también el nombre de reducción de la dimensión, ya que en general el conjunto de variables relevantes es inferior a las obtenidas

del bloque de preprocesado.

3. El último bloque realiza, a partir de las características adquiridas, el **aprendizaje** en sí. En problemas de clasificación, el vector de características ha de ser asignado a una de entre una serie de categorías. En problemas de regresión, se debe inferir el valor de una variable numérica continua. En problemas de agrupamiento, se asume que los datos pueden ser divididos en conjuntos diferenciados y de comportamiento homogéneo.

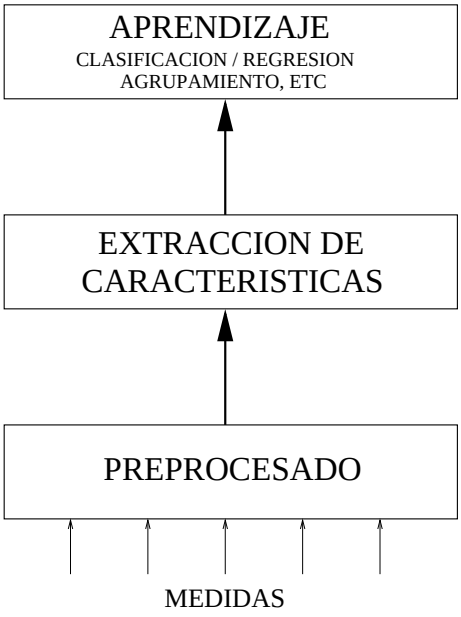

Figura 1.1: Esquema general de un sistema de aprendizaje basado en muestras.

La presente tesis aborda principalmente el problema de la extracción de características para clasificación. En este capítulo se realiza una revisión del estado del arte en métodos de reducción de la dimensión en el contexto de aprendizaje máquina. La amplitud de la bibliografía sobre reducción de la dimensión exige una revisión extensa, ya que el problema puede ser abordado de formas de diversa naturaleza. Además, el reciente interés mostrado en el ámbito del aprendizaje máquina hacia las técnicas basadas en teoría de la información requiere un análisis del estado del arte

para poder situar el trabajo presentado en esta tesis en el marco de estas aportaciones previas. Por estas razones, este capítulo introductorio ocupa una parte considerable de la tesis.

En el Apartado 1.1 se describe el problema general de la reducción de la dimensión, y su papel en un esquema general de aprendizaje basado en muestras; se discutirá la necesidad de dicha reducción y se revisarán los métodos más relevantes en los que la reducción de la dimensión supone el objetivo en sí mismo del aprendizaje. El Apartado 1.2 trata más directamente el problema de la extracción de características supervisada, es decir aquella en la que el objetivo es la obtención de un conjunto de variables idóneas para afrontar un problema de clasificación. En el Apartado 1.3 se introducen algunos principios básicos de teoría de la información; se discute la idoneidad de los criterios basados en esta teoría, y se realiza una amplia revisión de los métodos que la aplican a problemas de reconocimiento de patrones, prestando especial interés a aquellos que se centran en la reducción de la dimensión. Por último, en el Apartado 1.4 se enumeran los objetivos planteados en la tesis, una vez han sido expuestas las limitaciones de los métodos presentes en la literatura para resolver algunos problemas del ámbito del aprendizaje basado en teoría de la información.

1.1. El problema de la reducción de la dimensión

Las técnicas de aprendizaje habitualmente procesan conjuntos de N muestras o ejemplos de una variable D dimensional, que conforman una matriz de la forma $\mathbf{X} \equiv \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \}$, con $\mathbf{x}_i \in \mathcal{R}^D$. Distintos niveles de suposición sobre el origen de su muestreo sugerirán el uso de una u otra técnica para su procesado. Por ejemplo, se podrá asumir o no que hayan sido idénticamente distribuidas, dependiendo de si las $\mathbf{x}_i$ pueden o no considerarse realizaciones de una variable aleatoria generada de acuerdo con la función densidad de probabilidad (FDP) $p(\mathbf{x})$, sea ésta conocida o no. Igualmente, se podrá suponer independencia o no entre las muestras, en función de si la probabilidad con que ha sido generada la muestra $\mathbf{x}_i$ depende o no de la muestra

$\mathbf{x}_j$, $j \neq i$. Las técnicas de aprendizaje máquina presentes en la literatura suelen asumir que las muestras son independientes e idénticamente distribuidas (i.i.d.), lo cual está justificado en muchas de sus aplicaciones.

Los datos contenidos en la matriz $\mathbf{X}$ pueden estar o no acompañados por datos adicionales dados por una segunda matriz $\mathbf{Y} \equiv \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, $\mathbf{y}_i \in \mathcal{R}^r$. En este caso, el problema de aprendizaje consistirá en inferir el valor de $\mathbf{y}$ a partir del de $\mathbf{x}$, en cuyo caso hablamos de aprendizaje supervisado. Si esta variable auxiliar es unidimensional, y además adopta valores discretos, $y_i \in \{c_1, c_2, \ldots, c_{N_C}\}$, el problema de inferir su valor recibe el nombre de clasificación. Si y es continua, el problema se denomina regresión. El caso general de $\mathbf{y}$ multidimensional se denomina multirregresión. En cualquiera de los casos, el aprendizaje tiene lugar a partir de un conjunto de entrenamiento etiquetado, es decir, un conjunto de muestras para el cual conocemos los valores de $\mathbf{y}$. El conocimiento adquirido a partir de este conjunto muestral servirá para predecir el valor de $\mathbf{y}$ en posteriores muestras de test no etiquetadas. El trabajo desarrollado en esta tesis se centra en el problema de clasificación.

Con frecuencia, tanto el número de muestras N como el número de componentes (o dimensiones) D puede exigir una purga, tanto en el número de elementos como en sus componentes. En el primer caso, la reducción se deberá realizar de forma que las muestras que se preservan son suficientemente representativas como para que podamos inferir propiedades extrapolables a las descartadas. También pueden eliminarse muestras que podamos asumir que no son representativas, porque correspondan a medidas erróneas de $\mathbf{x}$. La detección y eliminación de estas muestras atípicas (*outliers*) ha sido estudiada para evitar el sesgo o deterioro en el proceso de aprendizaje [Blum, 1997]. También se puede perseguir el objetivo de encontrar el subconjunto de muestras más representativo para trazar la función de clasificación o regresión. Algunos esquemas de aprendizaje, como los basados en máquinas de vectores soporte (*Support Vector Machines*, SVM), realizan esa purga implícitamente, ya que hacen uso de un criterio de máximo margen, que queda finalmente definido a partir de las muestras cercanas a la frontera de discriminación [Scholkopf, 2002]. Aun así, el problema de

la reducción de la dimensión de los datos ha recibido mucha más atención en la literatura que el de la eliminación de muestras. Además, el problema de la selección o filtrado de muestras relevantes escapa al ámbito de esta tesis, pero puede consultarse una amplia revisión de métodos en [Blum, 1997].

1.1.1. Razones para reducir la dimensión

Los principales motivos para llevar a cabo una reducción en el número de variables de un conjunto muestral son los siguientes:

1. La capacidad de generalización de lo aprendido (es decir, la capacidad con que podremos aplicarlo a nuevas muestras) es mejor cuanto menor es la dimensión cuando se compara con el número de muestras. Una alta dimensión conduce al sobreajuste (*overfitting*) [Bishop, 1995], que constituye una adaptación excesiva del modelo de aprendizaje a los datos de entrenamiento, lo que impide que éste sea difícil de extrapolar a nuevos datos. La teoría del aprendizaje estadístico introduce el concepto de dimensión de Vapnik-Chervonenkis o dimensión VC para distinguir la dimensión efectiva (aquella en la que el clasificador es capaz de trazar funciones de discriminación con error nulo) de la dimensión real [Vapnik, 1998]. La cota sobre la diferencia entre el error de generalización y el de entrenamiento es del orden de $\sqrt{\frac{h}{N}\log(\frac{2N}{h})}$, siendo h la dimensión VC. Por ejemplo, para el caso de un clasificador lineal tenemos $h = D + 1$, que es el número de muestras que el clasificador puede separar independientemente de su valor y etiquetado. Vemos entonces que para este caso la cota disminuirá con D/N. De esta forma, la teoría del aprendizaje estadístico viene a justificar el hecho ya conocido en aprendizaje neuronal de que la capacidad de generalización de una máquina empeora con el cociente D/N. Dentro del aprendizaje con redes neuronales o de funciones de base radial, el teorema de Kolmogorov establece que podremos trazar una frontera que clasifique correctamente todas las muestras de entrenamiento sin más que poner suficientes

neuronas en la capa oculta [Bishop, 1995], lo que puede ser interpretado como una transformación a un espacio de mayor dimensión para obtener fronteras de clasificación que puedan separar totalmente las clases. Sin embargo, la utilización de un número menor de dichas neuronas lleva a soluciones con mayor capacidad de generalización.

2. El coste computacional de la tarea de aprendizaje disminuye con la dimensión de los datos a los que se aplica. Esta motivación es común a los métodos de selección de muestras como a los de extracción de características.

3. El hecho de contar con una representación de los datos en baja dimensión facilita la interpretación de los mismos así como su visualización.

4. Experimentos neurofisiológicos ponen de relieve un preprocesado, por parte del sistema nervioso de los animales, de la información adquirida por las células sonsoriales, lo que supone una reducción del volumen de información a enviar al cerebro [Atick, 1992]. El aprendizaje basado en redes neuronales se inspira en la naturaleza fisiológica del cerebro para construir algoritmos de aprendizaje; del mismo modo, las evidencias de compresión de la información sensorial sugieren la necesidad de realizar una reducción del número de variables a considerar en los problemas de reconocimiento de patrones.

1.1.2. Formas de reducción de la dimensión

La manera más sencilla de reducir la dimensión o número de componentes de los datos es seleccionar un subconjunto de los mismos. En tal caso, se persigue que el nuevo conjunto de d componentes ($d \leq D$) conserve la información relevante sobre la estructura de los datos (en el caso del aprendizaje no supervisado) o sobre su capacidad para inferir la variable auxiliar y (en el caso del aprendizaje supervisado). La selección de características busca el mejor subconjunto de variables de acuerdo con un determinado criterio. Algunos criterios exigen evaluar cada uno de

los $\binom{D}{d}$ posibles subconjuntos de variables para determinar el mejor de ellos. Este procedimiento es inviable por su complejidad. Al no ser aplicable a datos que tengan un número de dimensiones apreciable, se suele acudir a un método secuencial para construir el mejor subconjunto. Los métodos secuenciales más usados son la selección incremental (*forward selection*), la eliminación decremental (*backward elimination*) y procedimientos basados en el algoritmo *branch-and-bound* [Guyon, 2003].

Los métodos incrementales actúan de forma iterativa, añadiendo en cada paso la variable que, incorporada a las que ya han sido escogidas, proporciona el valor máximo del criterio considerado. El esquema básico se resume en la Figura 1.2. En el primer paso, se escoge la característica f_k que hace máximo el valor del criterio $C(f_k)$. En problemas de clasificación, $C(\cdot)$ viene dada por la precisión del clasificador sobre algún conjunto de validación. A partir de ahí, en cada iteración n, el conjunto óptimo de n características S_n estará formado por la unión del conjunto anterior S_{n-1} y la característica que, unida a éste, proporciona el mejor valor de $C(S_{n-1} + f_k)$.

$S_0 = [];$
$\hat{f}_k = \arg\max_f C(f_k);$
$S_1 = \{\hat{f}_k\}$
Para $n = 2$ hasta d
    $\hat{f}_k = \arg\max_f C(S_{n-1} + f_k);$
    $S_n = S_{n-1} + \hat{f}_k$
Fin

Figura 1.2: Esquema básico de selección de características secuencial incremental.

La eliminación decremental opera de forma inversa: se parte de los datos sin reducir e iterativamente se van eliminando variables. En cada caso se elimina la característica que menos degrada el valor del subconjunto resultante. El procedimiento se ilustra en la Figura 1.3. En este caso, partimos de un conjunto S_0 formado por todas las variables. En cada paso n, se elimina la característica cuya ausencia tiene

menor incidencia en el valor de $C(S_{n-1} - f_k)$.

$S_0 = [f_1, f_2, \ldots, f_D]$;
$\hat{f}_k = \arg\max_f C(S_0 - f_k)$;
$S_1 = \{S_0 - \hat{f}_k\}$ (Eliminamos la de menor incidencia sobre C)
Para $n = 2$ hasta d
    $\hat{f}_k = \arg\max_{f_k} C(S_{n-1} - f_k)$;
    $S_n = S_{n-1} - \hat{f}_k$ (Eliminamos la variable sin la cual
    el contraste es máximo)
Fin

Figura 1.3: Esquema básico de selección de características secuencial decremental.

El algoritmo *branch-and-bound* fue propuesto por primera vez en [Land, 1960] para resolver problemas de programación numérica con enteros, así como de optimización de problemas de complejidad no polinomial. Para que *branch-and-bound* sea aplicable al problema de selección de características, debe poder asumirse monotonicidad en las prestaciones con el número de variables. Es decir, un conjunto de m características debe ser, como mínimo, tan bueno como uno de $m - 1$ al que se ha añadido otra [Fukunaga, 1990]. Este algoritmo, aplicado al problema de la selección de características, funciona mediante la exploración del grafo que define las posibles secuencias de eliminación de variables, y que proporciona el valor del criterio $C(\cdot)$ en casa caso. La adecuada exploración de este grafo da lugar a un procedimiento eficiente de eliminación de variables.

La selección incremental es, de las mencionadas, la que mayor relevancia tiene en selección de características, ya que es el procedimiento seguido en una gran parte de los métodos propuestos en la literatura. Esto se debe a que es, en general, menos costoso en cuanto a computación que la eliminación secuencial: si el número de componentes a seleccionar es menor que la mitad del número de componentes iniciales, la búsqueda incremental necesita menos iteraciones que la decremental. Además la evaluación del coste $C(\cdot)$ es más rápido en el caso de la búsqueda incremental, ya que

se necesita menos cómputo cuando el número de variables es pequeño. El requisito de monotonicidad exigido a *branch-and-bound* es, en general, bastante fuerte, lo que dificulta su aplicación en muchos casos.

La extracción de características es un concepto más general que el de selección, ya que consiste en una función vectorial $\mathbf{f} : \mathcal{R}^D \rightarrow \mathcal{R}^d$. Si $\mathbf{f}$ viene dada por una transformación lineal, hablamos de extracción lineal, y puede expresarse como $\mathbf{f}(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$.

Es fácil ver que un método de selección puede ser expresado como una función de extracción lineal. En ese caso, $\mathbf{W}$ sería una matriz indicadora en la que cada columna constaría de D elementos, de los cuales todos serían nulos excepto el que selecciona la variable escogida, que valdría uno. Cada columna tendría un único elemento no nulo.

La literatura sobre aprendizaje ha abordado intensamente el tema de la selección de características. Sin embargo, la extracción ha sido menos tratada y, aunque existen métodos procedentes de la estadística clásica, constituye en general un problema de interés más reciente.

Los métodos de selección y de extracción de características supervisados se distinguen por el hecho de que el módulo de clasificación o regresión "supervisa" el proceso de extracción/selección de variables. Así, la función de coste $C(\cdot)$ presente en las Figuras 1.2 y 1.3 incorpora información sobre la capacidad de las variables obtenidas para inferir el valor y. En el caso no supervisado, el propio proceso de extracción de patrones puede ser considerado como el aprendizaje por sí mismo. Se espera que estas características en baja dimensión conserven la información relevante sobre la estructura intrínseca de los datos.

1.1.3. Reducción de la dimensión como aprendizaje no supervisado

En la literatura se ha descrito una gran cantidad de métodos de extracción de características que no son guiados por un método de clasificación o de regresión. En

estos casos, el objetivo de la reducción de la dimensión es extraer un conjunto de valores que sea de la mayor utilidad para la interpretación de los datos así como la identificación de la estructura o variedad conforme a la que están generados. Si esta variedad puede ser descrita en un espacio de dimensión menor a la original, el objeto de la reducción de la dimensión es proyectar los datos en este nuevo espacio. A continuación se describen brevemente algunos de los métodos más importantes, distinguiendo entre métodos lineales y no lineales.

Métodos lineales

La técnica lineal clásica por antonomasia es el Análisis de Componentes Principales (*Principal Component Analysis*, PCA) [Bishop, 1995] [Fukunaga, 1990]. Este método realiza un análisis espectral de la matriz de covarianza de $\mathbf{x}$, $\mathbf{C_x}$. Dicha matriz admite una descomposición en autovalores y autovectores de la forma $\lambda_k \mathbf{v}_k = \mathbf{C_x}\mathbf{v}_k$. Los autovalores λ_k son reales y mayores o iguales que cero, dado que toda matriz de covarianza es semidefinida positiva. Cada autovalor indica la dispersión o energía asociada a la dirección dada por su correspondiente autovector. Una proyección del tipo $\mathbf{z} = \boldsymbol{\lambda}^{-1/2}\mathbf{V}^T\mathbf{x}$, donde $\boldsymbol{\lambda} = \text{diag}(\{\lambda_1, \lambda_2, \cdots, \lambda_D\})$ y $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_D\}$, tiene la propiedad de proporcionar un conjunto de variables normalizadas y decorrelacionadas, es decir, con $\mathbf{C_z} = \mathbf{I}$. Si llevamos a cabo la proyección únicamente sobre los autovectores asociados a los mayores autovalores, proyectamos los datos sobre las direcciones de mayor varianza. El teorema de Karhunen-Loéve demuestra que PCA proporciona las proyecciones de menor error cuadrático medio. Es decir, la representación de un vector $\mathbf{x}$ sobre los d autovectores asociados a los autovalores más energéticos es la de menor error de reconstrucción de entre todas las trasformaciones que podríamos hacer a d dimensiones [Fukunaga, 1990]. Además de las direcciones de mayor energía, PCA sirve también para encontrar las de menor varianza, lo que también tiene gran importancia en procesado de señal. Algunas aplicaciones centran su atención en las proyecciones asociadas a los autovalores más pequeños, que determinan el denominado subespacio de ruido [Feng, 2005].

La versión probabilística de PCA presentada en [Tipping, 1999] utiliza un modelo de variables latentes para determinar la relación entre las observaciones y las proyecciones. Este modelo es de tipo lineal, por lo que forma parte de la familia de técnicas de análisis de factores y tiene la forma

$$\mathbf{x} = \mathbf{A}^T\mathbf{z} + \mathbf{m} + \boldsymbol{\epsilon}$$

donde $\mathbf{m}$ es la media que se introduce en el modelo cuando trabajamos con datos no centrados, y $\boldsymbol{\epsilon}$ es el error de reconstrucción, para el que supone estadística gausiana con varianza σ^2. De esta forma, el modelo probabilístico que relaciona a $\mathbf{x}$ y $\mathbf{z}$ es

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{A}^T\mathbf{z} + \mathbf{m}, \sigma^2\mathbf{I})$$

De esta forma, la matriz $\mathbf{A}$ es obtenida mediante un ajuste de máxima verosimilitud del modelo.

PCA es una técnica muy extendida cuando el criterio para la reducción de la dimensión no hace uso de momentos sobre los datos de orden mayor que dos; sin embargo, es insuficiente si pretendemos obtener un conjunto de variables que, además de decorrelacionadas, sean estadísticamente independientes. Para conseguir este nuevo objetivo surgen las técnicas, también lineales, de análisis de componentes independientes (*Independent Component Analysis*, ICA) [Hyvarinen, 2001]. En este caso, se asume que $\mathbf{x}$ ha sido generada a partir de la variable desconocida $\mathbf{s}$ mediante la transformación también desconocida $\mathbf{x} = \mathbf{A}^T\mathbf{s}$. La tarea consiste en, de forma ciega, obtener la inversión de la transformación $\mathbf{A}$.

Un esquema ICA básico consta de una primera fase en la que se aplica PCA y una segunda en la que, mediante una rotación o transformación ortonormal obtenemos la independencia entre señales como añadido a la decorrelación. Debido a que los esquemas ICA hacen uso de criterios basados en teoría de la información, serán tratados en el Apartado 1.3.4 con mayor detalle.

En algunas aplicaciones, la naturaleza no negativa del problema hace que la descomposición de un vector en sus componentes principales no refleje la naturaleza

aditiva de los datos. Por ejemplo, en tratamiento digital de imágenes, los valores de los píxeles son no negativos, y la imagen puede ser expresada como una composición de los elementos que la conforman. El método de factorización matricial no-negativa (*Non-negative Matrix Factorization*, NMF) proporciona una factorización aditiva, y constituye una interesante alternativa a PCA. Su utilidad ha sido demostrada en el aprendizaje no supervisado de imágenes faciales y categorización de documentos [Lee, 1999]. Las restricciones de no-negatividad fuerzan que un vector sólo pueda representarse en la base obtenida mediante adiciones de partes; es decir, los coeficientes que proyectan el vector en la base son siempre no negativos, por lo que la combinación lineal de los elementos de la base es estrictamente aditiva. PCA lleva a cabo la proyección $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ de forma que $\mathbf{W}$ puede ser una matriz de rango no completo, por lo que se puede llevar a cabo una reconstrucción con pérdidas $\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{z}$. En el caso de que $\mathbf{W}$ y $\mathbf{A}$ sean de rango completo, se cumple $\mathbf{W} = \mathbf{A}^{-1}$. Los vectores de la matriz $\mathbf{A}$ pueden ser interpretados como una base de representación, de forma que el vector $\mathbf{z}$ indica los coeficientes de representación en esa base. En el caso de NMF, se fuerza que las matrices $\mathbf{W}$ y $\mathbf{A}$ sean tales que el vector de reconstrucción $\mathbf{z}$ esté constituido por elementos no negativos. La matriz $\mathbf{A}$ y los vectores $\mathbf{z}_i$ se eligen conjuntamente para que maximicen la verosimilitud

$$L = \sum_{i=1}^{N} \left(\mathbf{z}_i^T \log \mathbf{A}^T\mathbf{z}_i - \mathbf{1}^T\mathbf{A}^T\mathbf{z}_i \right)$$

donde $\mathbf{1}$ es un vector de unos. Esta verosimilitud corresponde a un modelo probabilístico que se interpreta de forma que cada vector $\mathbf{x}_i$ es generado mediante un ruido de Poisson que se suma a la proyección $\mathbf{A}^T\mathbf{z}_i$.

Hasta el momento, se han descrito algunos de los métodos más relevantes de extracción de características lineal no supervisada. La principal limitación de estos métodos es su incapacidad de hacer frente a situaciones en que los datos se articulan conforme a una estructura subyacente que sólo puede ser descrita mediante proyecciones no lineales. En el siguiente apartado se describen las técnicas más relevantes de extracción no lineal de características.

Métodos no lineales

Dentro de las técnicas no lineales, las basadas en escalado multidimensional (*Multidimensional Scaling*, MDS) tratan de aplicar una transformación no lineal a los datos de forma que en el espacio d-dimensional resultante se conserven las distancias euclídeas entre muestras del espacio original, o al menos se introduzca la mínima distorsión [Backer, 1998]. El ejemplo clásico que ilustra el problema es el relativo a la cartografía: ¿qué proyección es la que mejor conserva las distancias al plasmar un mapa-mundi en dos dimensiones?

El concepto de preservación de distancias entre muestras (pero en este caso de forma local) es utilizado en el método de ajuste localmente lineal (*Locally Linear Embedding*, LLE) [Roweis, 2000]. LLE asume una estructura localmente lineal en los datos, de forma que cada muestra es susceptible de regresión lineal a partir de sus vecinos más próximos. Esta regresión se hace escogiendo los pesos w_{ij} que minimizan el error de reconstrucción local, de la forma

$$\arg\min_{\{w_{ij}\}} \sum_i |\mathbf{x}_i - \sum_j w_{ij}\mathbf{x}_j|^2 \tag{1.1}$$

donde w_{ij} es el coeficiente de regresión de la muestra $\mathbf{x}_i$ a partir de $\mathbf{x}_j$, con la restricción $\sum_j w_{ij} = 1$. Se pretende que la regresión tenga lugar sólo mediante las muestras vecinas, por lo que se establece la restricción adicional de que $w_{ij} = 0$ si $\mathbf{x}_j$ no forma parte de los K vecinos más próximos de $\mathbf{x}_i$ o bien no pertenece a la hiperesfera de radio r centrado en $\mathbf{x}_i$. Tanto K como r han de estar definidas de antemano. El funcional en (1.1) junto con estas restricciones da lugar a un problema de programación cuadrática, que permite obtener los w_{ij}. El siguiente paso es encontrar una transformación no lineal a una dimensión menor, de forma que el error de reconstrucción sea mínimo cuando se usan los coeficientes de reconstrucción que eran óptimos en el espacio inicial. Es decir, se buscan las proyecciones no lineales $\mathbf{z}_i$ que minimizan el contraste

$$\arg\min_{\{\mathbf{z}_i\}} \sum_i |\mathbf{z}_i - \sum_j w_{ij}\mathbf{z}_j|^2 \tag{1.2}$$

Para hacer que el problema de minimizar el funcional en (1.2) esté bien condicionado, se añaden las restricciones $\sum_i \mathbf{z}_i = \mathbf{0}$ y $\frac{1}{N}\mathbf{z}_i^T\mathbf{z}_i = \mathbf{I}$. De esta forma, el problema se resuelve, de nuevo, mediante programación cuadrática.

Otra técnica similar a LLE es la proyección isométrica de patrones (*Isometric Feature Mapping*, IsoMap) [Tenenbaum, 2000]. IsoMap también hace uso de un esquema basado en los K vecinos más próximos (*K-Nearest Neighbors*, KNN). En este caso, el objetivo es obtener la proyección en baja dimensión que permita que las distancias geodésicas entre pares de muestras se conserven. Para ello se asume que, a nivel de los vecinos más próximos, las distancia euclídea y geodésica coinciden, por lo que la distancia entre muestras lejanas ha de ser estimada como la composición de distancias entre las muestras intermedias. El algoritmo funciona como se muestra en la Figura 1.4.

Inicializar la matriz $\mathbf{D}$ de distancias, donde $D_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$
Imponer la condición $D_{ij} = \infty$ si $\mathbf{x}_i$ y $\mathbf{x}_j$ no son vecinos.
Para $k = 1$ hasta N
    Actualizo cada distancia $D_{ij} = \text{mín}(D_{ij}, D_{ik} + D_{kj})$;
Fin
Obtener la proyección en d-dimensiones minimizando $\|\tau(\mathbf{D}) - \tau(\mathbf{D}_z)\|^2$, donde $\mathbf{D}_z$ es la matriz de distancias euclídeas en d dimensiones

Figura 1.4: Funcionamiento del algoritmo IsoMap.

El operador $\tau(\mathbf{D})$ realiza un centrado de las matrices de distancias. El mínimo global de $\|\tau(\mathbf{D}) - \tau(\mathbf{D}_z)\|^2$ se alcanza con cada $\mathbf{z}_i$ obtenida mediante la proyección en los d autovectores asociados a los mayores autovalores de la matriz $\tau(\mathbf{D})$.

Un método que se encontraría a medio camino entre el aprendizaje de tipo neuronal y la extracción de características es el de los mapas autoorganizados (*Self-Organizing Maps*, SOM) [Kohonen, 2001]. Los SOM constituyen una familia de redes neuronales que proyectan las muestras en una malla de unidades o neuronas que organizan sus vecindades de acuerdo con la estructura intrínseca de los datos. Las

aplicaciones de los SOM son numerosas en aprendizaje no supervisado, aunque su utilidad en el campo de reducción de la dimensión es limitada, ya que los esquemas presentes en la literatura contemplan una dimensión de salida reducida. Esta está determinada por la forma en la que se construye la capa neuronal de salida, que habitualmente consiste en una malla bidimensional de neuronas.

El esfuerzo por tratar de extender la filosofía de PCA a un esquema no lineal, conjuntamente con el desarrollo de las técnicas de aprendizaje conocidas como métodos núcleo (*kernel*), ha dado lugar al método conocido como *Kernel*-PCA (KPCA) [Schölkopf, 1998]. Los métodos núcleo llevan a cabo los procesos de aprendizaje (clasificación, regresión, etc.) en un espacio de Hilbert de alta (o incluso infinita) dimensión. Un núcleo es una función que da una medida de similitud entre dos puntos, de la forma

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle \tag{1.3}$$

donde $\boldsymbol{\phi}(\mathbf{x})$ es la función de transformación a dicho espacio de Hilbert. El teorema de Mercer garantiza la existencia de la factorización en (1.3) bajo la condición de que la función núcleo sea semidefinida positiva [Schölkopf, 1998]. Esto permite trabajar explícitamente con la función núcleo en lugar de con $\boldsymbol{\phi}(\mathbf{x})$. La función núcleo más popular en la literatura es la basada en funciones de base radial (*Radial Basis Function*, RBF), que tiene la forma

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

KPCA resuelve el problema de autovalores

$$N\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}$$

donde $\mathbf{K}$ es la matriz Gram, dada por $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

La n-ésima componente principal queda determinada por el autovector $\mathbf{v}^n$ en el espacio de Hilbert, correspondiente al autovalor λ_n, y viene dada por

$$\langle \mathbf{v}^n, \boldsymbol{\phi}(\mathbf{x}) \rangle = \sum_{i=1}^{N} \alpha_i^n \langle \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x}_i) \rangle = \sum_{i=1}^{N} \alpha_i^n k(\mathbf{x}, \mathbf{x}_i) = \mathbf{K}\boldsymbol{\alpha}^n$$

por lo que la proyección no lineal está definida por el vector de coeficientes $\boldsymbol{\alpha}^n$.

Una diferencia notable entre KPCA y su versión lineal es que ahora el número de autovalores válidos (mayores que 0) puede ser, cómo máximo, igual al número de muestras N; en PCA este máximo era igual a D, la dimensión del espacio de representación.

También se ha definido una versión no lineal de ICA basada en núcleos. El método propuesto en [Bach, 2002] analiza la correlación entre dos o más proyecciones no lineales sobre las variables de entrada. Se demuestra que esta correlación es nula sólo si dichas proyecciones son estadísticamente independientes, en el caso del núcleo RBF.

1.2. Extracción de características supervisada

Hasta el momento se han revisado las técnicas que reducen la dimensión de forma no supervisada, obviando la labor de un bloque de procesado o aprendizaje posterior. En lo que está centrada la presente tesis es la extracción de características óptima para clasificación.

Existen dos metodologías diferenciadas para llevar a cabo la extracción de características. Según la aproximación guiada por clasificación (*wrapper*), los resultados de clasificación son realimentados al módulo de extracción o selección de características, de forma que el criterio que persigue el módulo de reducción de la dimensión es minimizar la probabilidad de error. Los métodos no guiados por la clasificación (*filter*), en cambio, siguen un criterio alternativo, de forma que no hay lazo de realimentación entre uno y otro bloque. Los métodos no guiados tratan de maximizar algún coste que esté relacionado con la capacidad de discriminación entre las clases; de esta forma, se persigue que el clasificador resuelva un problema más sencillo. En realidad, cabe distinguir una tercera categoría: los métodos incrustados (*embedded*), que llevan a cabo la selección de variables de forma acoplada al entrenamiento del clasificador [Guyon, 2003] . En lo sucesivo, los métodos incrustados se considerarán

como guiados por clasificación, ya que se encuentran conceptualmente más cerca de éstos.

1.2.1. Métodos guiados por clasificación

Los métodos *wrapper* recogidos en la literatura se han interesado mayoritariamente por la selección en lugar de la extracción de características. Esto se debe a que es más fácil determinar las variables que son o no relevantes para clasificación que construir una función de transformación sobre las variables que sea óptima en términos de error.

En [Kohavi, 1997] se propone un método de selección de tipo *wrapper* que es independiente del clasificador. Este método funciona construyendo un grafo en el que cada nodo representa a un subconjunto de variables, codificado mediante una secuencia binaria de longitud D, en la que cada bit indica si la variable ha sido seleccionada o no. Dos nodos del grafo están conectados si su distancia (de Hamming) es 1. El tamaño del grafo es 2^D. La búsqueda del nodo óptimo se realiza mediante un motor de búsqueda en grafos.

Sin embargo, la mayor parte de los métodos propuestos en la literatura enfocan el problema de la selección de variables teniendo en cuenta el clasificador concreto para el que se eligen. En [Leray, 1999] se recopilan algunas de las técnicas guiadas por clasificación en el ámbito de las redes neuronales, y se proporciona una clasificación de estos métodos de acuerdo con tres categorías: métodos de primer, de segundo y de tercer orden. La diferencia entre ellos responde a si se tienen en cuenta sólo los parámetros de la red (los pesos) o si también lo hacen sus derivadas en primer o segundo orden.

La aparición de las máquinas de vectores soporte (*Support Vector Machines*, SVM) ha supuesto la propuesta de interesantes métodos de tipo guiado para mejorar la capacidad de generalización de las máquinas obtenidas (es decir, la capacidad para clasificar correctamente nuevas muestras). La función de clasificación trazada

por una SVM tiene la forma

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b = \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{1.4}$$

donde $\mathbf{w}$ es un vector de proyección en el espacio de Hilbert, y por tanto de alta o infinita dimensión. Esta transformación se puede explicitar mediante el uso de una función núcleo $k(\mathbf{x}_i, \mathbf{x})$ conforme a (1.3). Los pesos α_i se obtienen mediante la minimización del funcional

$$L = \|\mathbf{w}\|^2 + C \sum_i \xi_i \tag{1.5}$$

donde cada ξ_i denota el coste de clasificación de la muestra i, de acuerdo con alguna de las funciones de coste propuestas [Scholkopf, 2002]. La norma cuadrática, para clasificadores no lineales, viene expresada en términos del núcleo de la forma

$$\|\mathbf{w}\|^2 = \sum_i \sum_j \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

El método de selección basado en la norma l_0 trata de aplicar un vector de pesos a las muestras de entrenamiento previamente a la construcción de la máquina [Weston, 2001a]. Se intenta minimizar la norma l_0 de este vector de pesos, es decir, el número de elementos no nulos. Dado que este problema tiene complejidad combinatoria, los autores proponen distintos problemas de optimización que minimizan otros tipos de coste alternativos sobre los pesos.

La teoría de aprendizaje estadístico incluye una cota superior al error de generalización que ha sido propuesta para selección de características en [Weston, 2001b] y aplicada a extracción de características [Leiva-Murillo, 2004]. La cota a la probabilidad de error viene dada por

$$p_e \leq \frac{1}{N} \|\mathbf{w}\|^2 R^2 = \frac{1}{N} \frac{R^2}{M^2} \tag{1.6}$$

donde R es el radio de la hiperesfera que circunscribe a los datos y $M = 1/\|\mathbf{w}\|^2$ el margen de clasificación. Ambas magnitudes son tomadas sobre el espacio de características transformado de Hilbert. El método propuesto en [Weston, 2001b] sigue

un procedimiento similar al de la minimización de la norma l_0 de [Weston, 2001a], aplicando un vector de pesos a las variables, que se harán nulos cuando esas variables sean irrelevantes. Este resultado se generalizó en [Leiva-Murillo, 2004] para extracción de características, de forma que la función R^2/M^2 quedaba caracterizada en función de una matriz de transformación $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, sobre la que se aplica una maximización mediante ascenso por el gradiente $\dfrac{\partial R^2/M^2}{\partial \mathbf{W}}$.

En [Rakotomamonjy, 2003] también se ha propuesto la maximización de la cota en (1.6), pero incluyendo información sobre la derivada respecto a los pesos, estableciendo un paralelismo con los métodos de primer orden que se describían en [Leray, 1999] para las redes neuronales.

La eliminación recursiva de características (*Recursive Feature Elimination*, RFE-SVM) propone un método en el que iterativamente se van eliminando características, a la vez que se actualiza una lista con una ordenación de las características por relevancia [Guyon, 2002]. El criterio para eliminar una variable es su incidencia en el margen, es decir, se elimina la variable cuya ausencia no altera el valor del nuevo margen.

Modificando el funcional en (1.5) y usando norma l_1 en lugar de norma l_2 para penalizar la complejidad del modelo, se llega a soluciones dispersas. Esta propiedad ha sido aplicada a modelos lineales en [Bi, 2003] para problemas de regresión, lo que da lugar a una selección de características implícita, ya que muchos de los pesos se hacen nulos. Combinando varios modelos lineales para varios subconjuntos de muestras, se llega a un subconjunto de variables relevantes. Con ellas, se construye una SVM no lineal, que será el clasificador finalmente considerado.

En este apartado se ha prestado una especial atención a los métodos guiados basados en métodos núcleo, ya que éstos son los que están recibiendo mayor atención en la literatura. Estos métodos tienen el inconveniente de requerir una alta complejidad computacional, ya que la valoración de cada conjunto de variables exige la construcción y evaluación de un clasificador. Para conjuntos de datos con muchas muestras o muchas variables, la utilización de estos métodos puede ser inviable. En

tales casos, puede ser conveniente usar un método de tipo *filter*. En el siguiente apartado se realiza una revisión de estos métodos que no exigen la construcción de un clasificador.

1.2.2. Métodos no guiados por clasificación

Los métodos guiados exigen la evaluación del clasificador sobre los datos transformados de acuerdo con la función de selección/extracción $\mathbf{f}(\mathbf{x})$. Esto puede ser computacionalmente costoso para algunos métodos de clasificación. Además, se pretende que el conjunto de características extraídas sea el más representativo independientemente del clasificador considerado. Por ello es conveniente el uso de una función de coste alternativa a la propia del clasificador. Los métodos no guiados hacen uso de funciones de contraste que reflejan la disparidad entre las clases, para de esta forma encontrar la transformación que dé lugar a características lo más discriminativas como sea posible.

La mayor parte de los métodos propuestos en la literatura sobre reducción de la dimensión abordan el problema de la extracción lineal de características o bien el de la selección, mientras que las transformaciones no lineales han recibido una atención menor. En realidad, la decisión sobre el uso de un extractor de características lineal o no lineal viene condicionada por el tipo de clasificador que le sucede. Un extractor no lineal debe ser capaz de caracterizar la variedad sobre la que se distribuyen los datos, y proyectarla de forma que un simple clasificador lineal pueda discriminar entre clases. Del mismo modo, un método de extracción lineal no sería capaz de extraer las características no lineales que pudiera haber en los datos, por lo que necesitaría un clasificador que sí pudiese trazar fronteras de tipo no lineal. Como ejemplo del potencial de esta última metodología, se han descrito resultados en los que las prestaciones de un sencillo clasificador KNN son drásticamente mejoradas mediante una apropiada transformación lineal, encontrando una adecuada métrica ponderada de la forma $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{Q} (\mathbf{x}_i - \mathbf{x}_j)$ [Goldberger, 2005]. Buscar la mejor matriz de ponderación $\mathbf{Q}$ equivale a buscar la mejor proyección del tipo $\mathbf{Ax}$ ya

que, si $\mathbf{Q} = \mathbf{A}^T\mathbf{A}$, entonces $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j)^T(\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j)$. Para encontrar la $\mathbf{A}$ óptima, se define una relación de vecindad "blanda" en términos probabilísticos: si P_{ij} es la probabilidad estimada de que la muestra $\mathbf{x}_i$ sea vecina de $\mathbf{x}_j$, el funcional a maximizar tiene la forma de estimación de la precisión de un clasificador KNN

$$F(\mathbf{A}) = \sum_i \sum_{j \in I_{y_i}} P_{ij} = \sum_i \sum_{j \in I_{y_i}} \frac{\exp(-|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j|^2)}{\sum_{t \neq i} \exp(-|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_t|^2)}$$

donde I_{y_i} es el conjunto de índices de las muestras pertenecientes a la clase y_i. Dado que el funcional es fácilmente derivable respecto a $\mathbf{A}$, admite una maximización mediante ascenso por gradiente.

Un método de selección de tipo no guiado suele actuar en dos direcciones simultáneas a la hora de realizar la purga y ordenación de variables. En primer lugar, busca aquellas que son más relevantes para clasificación. En segundo lugar, trata de minimizar la redundancia entre las variables seleccionadas [Yu, 2004]. A continuación se realiza una revisión de los criterios de relevancia y de redundancia y los métodos propuestos para la maximización de aquella y la minimización de ésta.

Criterios de relevancia basados en contrastes lineales

En el ámbito de la clasificación, la relevancia de una característica viene dada por su capacidad de discriminación entre clases. Las variables que se han de buscar son por tanto las que maximicen el grado de separación entre las muestras correspondientes a las distintas clases.

El análisis de correlaciones canónicas (*Canonical Correlation Analysis*, CCA), realiza un análisis espectral similar al de PCA, pero ahora teniendo en cuenta la correlación cruzada entre dos variables, lo que nos permite llevar a cabo algún tipo de supervisión [Hotelling, 1936] [Hardoon, 2004]. El objetivo de CCA es obtener un par de transformaciones lineales tales que aplicadas a dos vectores de señales, la correlación entre las salidas es máxima. Para poder aplicar CCA a un problema de clasificación se debe acondicionar la variable que define a las clases a tal efecto. Para ello transformamos la variable y en el vector $\mathbf{y}$ mediante su dicotomización: $\mathbf{y}$ es un

vector de longitud N_c, y su posición j-ésima vale 1 si la muestra pertenece a la clase c_j; en otro caso vale 0. La correlación entre $\mathbf{W}_x^T\mathbf{x}$ y $\mathbf{W}_y^T\mathbf{y}$, que es lo que se pretende maximizar, viene dada por

$$\rho_{xy} = \frac{\mathbf{W}_x^T\mathbf{C}_{xy}\mathbf{W}_y}{(\mathbf{W}_x^T\mathbf{C}_{xx}\mathbf{W}_x)^{1/2}(\mathbf{W}_y^T\mathbf{C}_{yy}\mathbf{W}_y)^{1/2}} \tag{1.7}$$

El problema se puede plantear como una descomposición generalizada en autovalores y autovectores de la forma $\mathbf{Aw} = \lambda\mathbf{Bw}$, donde $\mathbf{A} = \begin{pmatrix} \mathbf{C}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_y \end{pmatrix}$ y $\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{C_{xy}} \\ \mathbf{C_{yx}} & \mathbf{0} \end{pmatrix}$.

Las proyecciones que maximizan la correlación en (1.7) son las dadas por los autovectores asociados a los mayores autovalores.

El análisis de discriminantes lineales (*Linear Discriminant Analysis*, LDA) fue propuesto por Fisher en los años 30, y aún proporciona prestaciones competitivas cuando es comparado con métodos del estado del arte [Fukunaga, 1990]. En LDA, se ha de maximizar una magnitud que exprese la distancia entre las clases, a la par que minimizar un valor que exprese la dispersión de las mismas [Li, 2006]. El esquema clásico busca la matriz de proyección $\mathbf{W}$ que maximiza la siguiente magnitud

$$J_F(\mathbf{W}) = \mathrm{tr}\Big(\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_w\mathbf{W}}\Big) \tag{1.8}$$

donde $\mathbf{S}_b = \sum_{l=1}^{N_c} P(c_l)(\mathbf{m}_l - \mathbf{m})(\mathbf{m}_l - \mathbf{m})^T$ es la matriz de dispersión entre clases, siendo $\mathbf{m}_l$ la media de la clase c_l, $\mathbf{m}$ la media de todas las muestras y $P(c_l)$ la probabilidad a-priori de la clase c_l; $\mathbf{S}_w = \sum_{l=1}^{N_c} P(c_l)\mathbf{C}_l$ es la matriz de dispersión interna, siendo $\mathbf{C}_l$ la matriz de covarianza de la clase c_l. La solución de este problema viene dada por los autovectores de la matriz $\mathbf{S}_w^{-1}\mathbf{S}_b$. Los autovectores asociados a los mayores autovalores son los que proporcionan el valor más alto del contraste. El coste de LDA en (1.8) contempla un doble objetivo. Por un lado, minimiza la dispersión entre las muestras pertenecientes a una misma clase. Por otro, maximiza la dispersión entre las muestras que pertenecen a distintas clases.

El esquema básico LDA presenta varios inconvenientes. Para empezar, el rango de la matriz $\mathbf{S}_w^{-1}\mathbf{S}_b$ es $N_c - 1$, por lo que éste es el número máximo de autovalores positivos resultantes y, por tanto, la dimensión máxima a la salida. En un problema bi-clase, proporciona una única proyección, que puede considerarse como la normal del hiperplano de separación entre las clases.

Otro problema de LDA (aunque no es exclusivo a este método) es la degeneración de la solución cuando hay pocas muestras en comparación con la dimensión (lo que se denomina en la literatura "*small sample size problem*"). En ese caso, la matriz $\mathbf{S}_w$ puede tener aun menos autovalores positivos. Se han propuesto muchos esquemas para aliviar este efecto, básicamente tratando de regularizar la matriz $\mathbf{S}_w$. Una revisión de los métodos propuestos para este fin puede encontrarse en [Li, 2006].

Para paliar problemas de bajo tamaño muestral se ha planteado otra alternativa a LDA que se basa en un criterio de máximo margen [Li, 2006]. El criterio de dispersión tiene ahora la forma

$$J'_F = \frac{1}{2}\sum_{l=1}^{N_c}\sum_{j=1}^{N_c} P(c_l)P(c_j)d(c_l, c_j)$$

donde $d(\cdot,\cdot)$ es alguna medida de distancia entre clases. Para tener en cuenta tanto la dispersión entre clases como la interna, se propone el uso de la distancia

$$d(c_l, c_j) = d_E(\mathbf{m}_l, \mathbf{m}_j) - \mathrm{tr}(\mathbf{C}_l) - \mathrm{tr}(\mathbf{C}_j)$$

donde $d_E(\mathbf{m}_l, \mathbf{m}_j) = (\mathbf{m}_l - \mathbf{m}_j)(\mathbf{m}_l - \mathbf{m}_j)^T$, y los otros dos términos penalizan la dispersión interna de las clases.

En [Martínez, 2005] se propone un análisis de las matrices $\mathbf{S}_w$ y $\mathbf{S}_b$ para concluir en qué casos LDA proporciona proyecciones fiables (y por tanto cuándo es conveniente su aplicación) y en cuáles no, en términos de ángulos entre los autovectores de ambas matrices. Para aliviar el problema que supone que este ángulo sea pequeño (lo que se traduce en una baja capacidad de generalización) se sugiere realizar la búsqueda de las proyecciones en el subespacio complementario a los autovectores de $\mathbf{S}_w$.

Otra limitación de LDA es el hecho de no conservar las distancias entre las clases en el espacio proyectado, a no ser que se tomen los autovectores de todos los au-

tovalores positivos. Para contrarrestar este problema, se ha propuesto el algoritmo basado en el criterio de Fisher ponderado "por parejas" (*Weighted Pairwise Fisher Criteria*, aPAC) [Loog, 2001]. La definición alternativa para $\mathbf{S}_b$ en aPAC es

$$\mathbf{S}_B = \sum_{l=1}^{N_c-1} \sum_{j=l+1}^{N_c} P(c_l)P(c_j)(\mathbf{m}_l - \mathbf{m}_j)(\mathbf{m}_l - \mathbf{m}_j)^T \tag{1.9}$$

Esta nueva matriz tiene la propiedad de considerar distancias entre clases una a una, lo que permite tener en cuenta la particularidad de cada par de clases.

Otra crítica que se ha hecho a LDA es su incapacidad para tratar correctamente los datos heteroscedásticos (es decir, aquellos distribuidos conforme a matrices de covarianza muy distintas para cada clase). Aunque el contraste a maximizar por LDA (1.8) tiene en cuenta dichas matrices en la matriz de dispersión interna $\mathbf{S}_w$, no lo hace en la matriz de dispersión externa $\mathbf{S}_b$, perdiéndose una información valiosa sobre la disparidad de estructura de los datos [Loog, 2004]. Para paliar esta limitación se propone el uso de la distancia de Chernoff como criterio de separabilidad [Fukunaga, 1990]. La distancia de Chernoff entre dos clases c_l y c_j está definida a partir de sus funciones de densidad de la forma

$$d_C(c_l, c_j) = -\log \int p^{\alpha}(\mathbf{x}|c_l)p^{1-\alpha}(\mathbf{x}|c_j)d\mathbf{x} \tag{1.10}$$

donde $\alpha \in (0, 1)$, y $p(\mathbf{x}|c_l)$ es la densidad de probabilidad de $\mathbf{x}$ condicionada a la clase c_l. Aunque en [Cover, 2006] se define la distancia de Chernoff como la menor de todas las proporcionadas barriendo el valor de α, en este trabajo se ajusta, como se verá, mediante otro criterio. La distancia de Chernoff tiene la propiedad de que su particularización para distribuciones normales es resoluble analíticamente y da lugar a un discriminante en términos de momentos de primer y segundo orden

$$\begin{aligned} d_C(c_l, c_j) = (\mathbf{m}_l - \mathbf{m}_j)^T \big(\alpha\mathbf{C}_l + (1-\alpha)\mathbf{C}_j\big)^{-1}(\mathbf{m}_l - \mathbf{m}_j) \\ + \frac{1}{\alpha(1-\alpha)} \log \frac{|\alpha\mathbf{C}_l + (1-\alpha)\mathbf{C}_j|}{|\mathbf{C}_l|^{\alpha}|\mathbf{C}_j|^{1-\alpha}} \end{aligned} \tag{1.11}$$

Basándose en esta distancia, en [Loog, 2004] se propone la siguiente definición alter-

nativa de la matriz de dispersión $\mathbf{S}_b$

$$\mathbf{S}_b = \mathbf{S}^{-1/2}(\mathbf{m}_l - \mathbf{m}_j)(\mathbf{m}_l - \mathbf{m}_j)^T\mathbf{S}^{-1/2} + \frac{1}{\alpha(1-\alpha)}(\log \mathbf{S} - \alpha \log \mathbf{C}_l - (1-\alpha)\mathbf{C}_j) \quad (1.12)$$

donde $\mathbf{S} = \alpha\mathbf{C}_l + (1-\alpha)\mathbf{C}_j$.

El ajuste del valor de α se realiza teniendo en cuenta que la matriz $\mathbf{S}_w$ según (1.12) y la que se definió para LDA en (1.8) han de coincidir para el caso homoscedástico (idénticas matrices de covarianza para cada clase). Eso lleva a establecer los valores $\alpha = P(c_0)$ y $1 - \alpha = P(c_1)$ para el caso binario. La generalización al caso multiclase es inmediato.

La particularización de (1.10) al caso $\alpha = 1/2$ da lugar a la distancia de Bhattacharyya [Fukunaga, 1990], que también ha sido usada como criterio en selección y extracción de características. Esta distancia se deduce fácilmente de (1.11) y viene dada por

$$d_B(c_l, c_j) = \frac{1}{8}(\mathbf{m}_l - \mathbf{m}_j)^T \left(\frac{\mathbf{C}_l + \mathbf{C}_j}{2}\right)^{-1} (\mathbf{m}_l - \mathbf{m}_j) + \frac{1}{2}\log \frac{|(\mathbf{C}_l + \mathbf{C}_j)/2|}{|\mathbf{C}_l|^{1/2}|\mathbf{C}_j|^{1/2}} \quad (1.13)$$

La distancia de Bhattacharyya proporciona una cota superior al error de Bayes [Fukunaga, 1990]. La cota viene dada por

$$p_e \leq \frac{1}{2}\exp(-d_B)$$

donde p_e es el error de Bayes. Aunque la cota es poco ajustada, legitima el uso de d_B como criterio para extracción de características [Hsieh, 2006].

El uso de la distancia de Bhattacharyya ha sido propuesto para selección de características en [Mao, 2003]. En este caso, se lleva a cabo una ortogonalización de los datos que da lugar a matrices de covarianza diagonales. De esta forma, se puede evaluar d_B para cada característica individualmente. Mediante una búsqueda incremental, se añaden una a una las variables asociadas a una mayor distancia en cada paso.

Otro esquema basado en la distancia de Bhattacharyya es el que se propone en [Hsieh, 2006]. Este trabajo está basado en un algoritmo propuesto en [Hsieh,

1998], que obtiene características relevantes alternando el criterio de disparidad de medias y el de disparidad de covarianza. El procedimiento consiste en combinar el criterio aPAC ya explicado y la extracción de características con medias comunes (*Common-Mean Feature Extraction*, CMFE). Obteniendo secuencialmente proyecciones de acuerdo con ambos criterios se logran proyecciones que sean discriminativas tanto en distancia entre sus medias como en disparidad en sus varianzas. Paralelamente a este procedimiento, se proporciona una estimación de la probabilidad de error. Para ello se propone una distancia construida de forma heurística a partir de simulaciones de forma que la cota de Bhattacharyya pasa a convertirse en una estimación. Es decir, se propone una distancia d'_B, similar a la mostrada en (1.13) pero con los coeficientes y exponentes modificados de forma que la aproximación $p_e \approx -0{,}5\exp(d'_B)$ sea válida.

Una forma razonable de llevar a cabo selección de características sobre clasificadores lineales es diseñar las funciones de discriminación del tipo $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ para interpretar el resultado en términos de selección y valoración de la importancia de cada característica: cada elemento de $\mathbf{w}$ proporciona una idea del peso que tiene cada variable en la función de discriminación. Una generalización de la desigualdad de Chebychev-Cantelli proporciona una cota al error de clasificación [Bhattacharyya, 2004] de estos discriminantes. Si llamamos $\mathcal{H}$ al hiperplano que cumple $\mathbf{w}^T\mathbf{x} < b$, la desigualdad se expresa como

$$P(\mathbf{x} \in \mathcal{H}) \geq \frac{s^2}{s^2 + \mathbf{w}^T\mathbf{C}\mathbf{w}}$$

donde $s = \text{máx}(b - \mathbf{w}^T\mathbf{m}, 0)$, siendo $\mathbf{m}$ y $\mathbf{C}$ la media y covarianza de $\mathbf{x}$. Una formulación basada en programación cuadrática con restricciones cónicas da lugar al algoritmo propuesto en [Bhattacharyya, 2004].

Para finalizar, se revisan algunos de los métodos de extracción basados en discriminación lineal y que son considerados como métodos clásicos en la literatura. Los más relevantes son los siguientes:

Regresion inversa "en rodajas" (*Sliced Inverse Regression*, SIR). Este método

realiza la extracción de características mediante una regresión inversa, en la que $\mathbf{x}$ es estimada a partir de y. El dominio de y ha de ser discretizado, por lo que es dividido en "rodajas". En un problema de clasificación, esta cuantificación ya está hecha, y hay tantas rodajas como clases. Entonces se lleva a cabo un procedimiento de tipo PCA a una matriz de covarianza ponderada $\mathbf{M} = \sum_{k=1}^{N_c} P(c_k)\mathbf{m}_k\mathbf{m}_k^T$. La matriz de transformación $\mathbf{W}$ se obtiene de los autovectores asociados a los autovalores mayores de $\mathbf{M}$ [Li, 1991].

Mínimos cuadrados parciales (*Partial Least Square*, PLS). Se considera una versión supervisada de PCA, dado que tiene en cuenta los coeficientes de error mínimo-cuadrático que permiten hacer regresión lineal sobre y [Hastie, 1998]. Se parte de una proyección que aproxima y a partir de $\mathbf{x}$ con un error mínimo; es decir, se obtiene $\mathbf{w}_1$ tal que $t = \mathbf{w_1}^T\mathbf{x}$ es una estimación de y de mínimo error. El resto de las proyecciones se obtienen de la misma forma, pero buscando en el espacio complementario a las proyecciones ya obtenidas.

Direcciones principales de la Hessiana (*Principal Hessian Directions*, PHD). Este método trata de encontrar las componentes principales de la matriz Hessiana de la función de reconstrucción de y a partir de $\mathbf{x}$. Si se asume linealidad en esta función, dicha matriz está constituida por coeficientes de regresión lineales. El objetivo es encontrar las direcciones principales de esta matriz de coeficientes, lo que lleva de nuevo a un problema de descomposición en autovalores y autovectores [Li, 1992].

Criterios de redundancia entre variables

En la mayoría de los métodos extracción de características propuestos en la literatura, una forma de garantizar una baja redundancia en las proyecciones obtenidas es imponer ortogonalidad en dichas proyecciones. De esta forma se puede inducir decorrelación sobre las características obtenidas, si las originales están decorrelacionadas. Aunque el hecho de que las proyecciones estén decorrelacionadas no significa

que sean estadísticamente independientes, la anulación de los momentos cruzados de primer y segundo orden puede eliminar gran parte de la dependencia estadística entre las características.

En los métodos de selección de características no se puede imponer ortogonalidad ni decorrelación. Algunos métodos de selección usan el concepto de "cubierta" de Markov (*Markov Blanket*, MB) para el análisis de la redundancia entre variables [Koller, 1996] [Fleuret, 2004], a partir de un modelo markoviano de relación entre las mismas. Es decir, se asume una dependencia del tipo $A \rightarrow B \rightarrow C$, de modo que $P(C|B,A) = P(C|B)$ y $P(C|A) = P(C)$. Se dice que un conjunto de variables M es una cubierta de otra variable f_k si el conjunto de todas las demás variables es condicionalmente independiente de f_k dado M; es decir, si

$$P(f_k|M, F - M - f_k) = P(f_k|M)$$

donde F es el conjunto de todas las variables. La cubierta de una variable, por tanto, es el conjunto de características tales que la probabilidad de aquella condicionada a éste no depende del resto de variables.

Con conjuntos de datos dados por variables discretas, las probabilidades pueden ser estimadas de forma sencilla a partir de las tablas empíricas de coocurrencias. Si las variables son continuas, se ha de estimar la densidad de probabilidad, lo que hace más difícil caracterizar estas relaciones de dependencia. En el siguiente apartado, se realiza una revisión de los métodos basados en teoría de la información, que hacen un uso intensivo del modelado de densidades de probabilidad, y que establece un marco más apropiado para el análisis conjunto de la relevancia y la redundancia presentes en un conjunto de variables.

1.3. Criterios basados en teoría de la información

La aparición de la teoría de la información de Shannon a mediados del siglo XX vino a cuantificar los conceptos de información o incertidumbre, así como a describir de forma matemática la transmisión de información a través de un canal de comunicaciones [Shannon, 1948]. Aparte de las telecomunicaciones, esta teoría ha tenido una profunda influencia en estadística, economía, etc.

Poco después de la formulación de esta teoría por Shannon, surgió la intuición de que los conceptos ahí descritos podrían explicar el procesado neurológico de la información sensorial en los humanos y animales. Ya el trabajo pionero de Attneave (1954) sugería que nuestro sistema perceptual elimina redundancias en su estimulación y codifica dicha información. Posteriores experimentos sobre la visión humana sugieren que nuestra capacidad perceptiva se sitúa en torno a los 40-50 bits/s aun cuando la tasa de adquisición de la retina llega hasta los 500.000 bits/s [Atick, 1992]. Experimentos similares con la capacidad auditiva arrojan resultados parecidos: la entropía perceptual de la música digital está cifrada en 2,1 bits/muestra, lo cual constituye la cota que tratan de alcanzar los sistemas de codificación perceptual de audio [Painter, 2000]. Este "cuello de botella" en cuanto a la información transmitida de los sensores al cerebro ha inspirado la búsqueda de esquemas basados en redes neuronales artificiales que reduzca la redundancia tanto como sea posible para mejorar las prestaciones de la misma.

En el campo de la reducción de la dimensión, por tanto, la teoría de la información ofrece un soporte indiscutible. Dado que el problema a resolver en extracción y selección de características es obtener características relevantes así como poco redundantes entre sí, esta teoría dispone de las herramientas necesarias para cuantificar el grado de dependencia entre cada característica a obtener y la variable de referencia (las clases) así como entre las propias características.

A continuación se expondrán los conceptos básicos de teoría de la información necesarios para su aplicación al problema del aprendizaje. Después se describe el

soporte teórico que justifica la aplicación de la información mutua como criterio en aprendizaje y reducción de la dimensión. Tras ello, se aborda el problema del modelado de FDPs, que es un paso previo en la mayoría de los métodos propuestos en la literatura que hacen uso de la teoría de la información. Por último se describe el estado del arte en este ámbito del aprendizaje, prestando especial atención a los métodos de reducción de la dimensión propuestos en la literatura.

1.3.1. Conceptos básicos de teoría de la información

La teoría de la información define el grado de dependencia estadística entre dos variables aleatorias a partir del concepto de información mutua. El nombre hace referencia al hecho de que lo que se mide es la información que contiene una variable acerca del valor de alguna otra. La información mutua puede definirse a partir de la divergencia de Kullback-Leibler (KL) [Cover, 2006]. Esta divergencia viene dada, para dos distribuciones de probabilidades discretas P y Q con el mismo dominio $X, Y \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\}$, por

$$D_{KL}(P,Q) = E_{P(\mathbf{x})}\left\{\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right\} = \sum_{l=1}^{L} P(\mathbf{x}_l)\log\frac{P(\mathbf{x}_l)}{Q(\mathbf{x}_l)}$$

En concreto, la información mutua entre las variables X e Y se define como la divergencia KL entre la distribución conjunta $P_{XY}(\mathbf{x},\mathbf{y})$ y el producto de las distribuciones marginales $P_X(\mathbf{x})$ y $P_Y(\mathbf{y})$

$$I(X,Y) = D_{KL}(P_{XY}, P_X P_Y) = \sum_{l_x=1}^{L_X}\sum_{l_y=1}^{L_Y} P_{XY}(\mathbf{x}_{l_x},\mathbf{y}_{l_y})\log\frac{P_{XY}(\mathbf{x}_{l_x},\mathbf{y}_{l_y})}{P_X(\mathbf{x}_{l_x})P_Y(\mathbf{y}_{l_y})}$$

ya que ahora X e Y pueden tener distinto dominio, $X \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{L_X}\}$, $Y \in \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{L_Y}\}$. La información mutua mide el grado de dependencia estadística entre dos o más variables. Es por tanto una medida de reducción de la incertidumbre que tenemos sobre el valor de una variable una vez que conocemos el de la otra.

El concepto de entropía o "autoinformación" se define sobre una única variable, cuando ésta es discreta, como

$$H(X) = I(X, X) = -\sum_{l=1}^{L} P_X(\mathbf{x}_l) \log P_X(\mathbf{x}_l)$$

La entropía recibe también el nombre de información media, ya que proporciona el valor medio de la información contenida por cada estado de una variable aleatoria, y que de acuerdo con la definición de Hartley y adoptada por Shannon [Arndt, 2001], es $I_{x_l} = \log \frac{1}{P(\mathbf{x}_l)}$.

De igual modo, la información mutua entre dos variables continuas se define a partir del operador integral en lugar del sumatorio, y a partir de las funciones de densidad de probabilidad (FDP) $p_X(\mathbf{x})$, $p_Y(\mathbf{y})$ y $p_{XY}(\mathbf{x}, \mathbf{y})$ en lugar de las funciones de distribución, de la forma

$$I(X, Y) = D_{KL}\left(p_{XY}(\mathbf{x}, \mathbf{y}), p_X(\mathbf{x}) p_Y(\mathbf{y})\right) = \int_{\mathbf{x}} \int_{\mathbf{y}} p_{XY}(\mathbf{x}, \mathbf{y}) \log \frac{p_{XY}(\mathbf{x}, \mathbf{y})}{p_X(\mathbf{x}) p_Y(\mathbf{y})} d\mathbf{x} d\mathbf{y} \tag{1.14}$$

La entropía diferencial es la generalización del concepto de entropía para variables aleatorias continuas, y se define como

$$h(X) = -\int p_X(\mathbf{x}) \log p_X(\mathbf{x}) d\mathbf{x} \tag{1.15}$$

Información mutua y entropía diferencial están relacionadas de la forma

$$I(X, Y) = h(X) - h(X|Y) \tag{1.16}$$

Esta tesis se centra en los problemas de clasificación, por lo que en lo sucesivo se asumirá una variable X multidimensional y continua y una variable Y discreta y unidimensional que puede tomar los valores $Y \in \{c_1, c_2, \ldots, c_L\}$. En tal caso, la integral en (1.14) para Y pasaría a ser un sumatorio

$$I(X, Y) = \sum_{l=1}^{L} \int_{\mathbf{x}} p_{XY}(\mathbf{x}, c_l) \log \frac{p_{XY}(\mathbf{x}, c_l)}{p_X(\mathbf{x}) P_Y(c_l)} d\mathbf{x} \tag{1.17}$$

Una definición de entropía alternativa a la de Shannon es la propuesta por Renyi [Arndt, 2001]. En realidad, Renyi define una familia de entropías que comprende a la propia entropía de Shannon. La entropía de Renyi de orden α se define como

$$h_R^\alpha(X) = \frac{1}{1-\alpha} \log \int p_X^\alpha(\mathbf{x}) d\mathbf{x} \tag{1.18}$$

Aunque el valor de (1.18) está indefinido para $\alpha = 1$, existe una equivalencia con la entropía de Shannon, ya que es fácil comprobar que $\lim_{\alpha \to 1} h_R^\alpha(X) = h(X)$.

De forma similar, Renyi redefine la divergencia entre FDPs, de la forma

$$D_R^\alpha(p, q) = \frac{1}{1-\alpha} \log \int \frac{p^\alpha(\mathbf{x})}{q^{\alpha-1}(\mathbf{x})} d\mathbf{x}$$

El caso particular de $\alpha = 2$ es el más extendido en la literatura y da lugar a la siguiente definición de información mutua

$$I_R(X, Y) = -\log \sum_k \int \frac{p_{XY}^2(\mathbf{x}, c_l)}{p_X(\mathbf{x}) P_Y(c_l)} d\mathbf{z}$$

La principal ventaja de la entropía y la información mutua de Renyi frente a las definiciones de Shannon es que la integral está dentro del logaritmo, lo que la hace más fácilmente tratable.

Aparte de la entropía de Renyi, se han definido otras divergencias alternativas que han dado lugar a distintas medidas de "pseudo"-información mutua. Éstas son escogidas de forma que cumplan las principales propiedades de la divergencia KL, es decir:

1. $D_{KL}(p, q) \geq 0$
2. $D_{KL}(p, q) = 0 \iff p = q$

En la geometría euclídea podemos encontrar propiedades que nos sirvan para definir estas medidas de divergencia. Uno de los resultados fundamentales de la geometría es que para dos vectores cualesquiera $\mathbf{u}$ y $\mathbf{v}$ se cumple

$$\|\mathbf{u} - \mathbf{v}\|^2 \geq 0 \tag{1.19}$$

Como puede verse, la magnitud $\|\mathbf{u}-\mathbf{v}\|^2$ cumple para los vectores las mismas propiedades que deseamos para una medida de divergencia sobre FDPs. Consideremos el espacio de las densidades de probabilidad como un espacio vectorial con el siguiente producto escalar

$$p \cdot q = \langle p, q \rangle = \int p(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}$$

Entonces, la desigualdad en (1.19) puede ser utilizada como medida de divergencia sin más que sustituir los vectores $\mathbf{u}$ y $\mathbf{v}$ por funciones de densidad sobre los que opera el producto escalar en (1.19). De esta forma definimos la divergencia cuadrática como

$$\begin{aligned} D_Q(p, q) &= \|p - q\|^2 = p \cdot p + q \cdot q - 2p \cdot q \\ &= \int p^2(\mathbf{v}) d\mathbf{v} + \int q^2(\mathbf{v}) d\mathbf{v} - 2 \int p(\mathbf{v}) q(\mathbf{v}) d\mathbf{v} \end{aligned} \tag{1.20}$$

Puede verificarse fácilmente que esta divergencia cumple las dos propiedades mencionadas de D_{KL}.

Por otro lado, la desigualdad de Cauchy-Schwartz nos dice

$$\|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \geq (\mathbf{u}^T \mathbf{v})^2 \tag{1.21}$$

que también puede plantearse en términos de una magnitud que sea no negativa

$$\frac{\|\mathbf{u}\|^2 \|\mathbf{v}\|^2}{(\mathbf{u} \cdot \mathbf{v})^2} \geq 1 \rightarrow \log \frac{\|\mathbf{u}\|^2 \|\mathbf{v}\|^2}{(\mathbf{u} \cdot \mathbf{v})^2} \geq 0$$

Igual que en el caso de la desigualdad cuadrática en (1.19), queda traducido en términos de divergencia de FDPs como

$$\begin{aligned} D_{CS}(p|q) &= \log \frac{\|p\|^2 \|q\|^2}{(p \cdot q)^2} \\ &= \log \int p^2(\mathbf{v}) d\mathbf{v} + \log \int q^2(\mathbf{v}) d\mathbf{v} - 2 \log \int p(\mathbf{v}) q(\mathbf{v}) d\mathbf{v} \end{aligned} \tag{1.22}$$

Del mismo modo que D_{KL}, definida sobre las densidades conjunta y marginal, da lugar a la definición de información mutua, el uso de las divergencias en (1.20) y (1.22) da lugar a definiciones de pseudo-información mutua ampliamente utilizadas en aprendizaje basado en teoría de la información.

1.3.2. La información mutua como criterio para reducir la dimensión

En lo sucesivo, y por simplicidad de notación, consideraremos la variable $\mathbf{x}$ en el espacio de características original D-dimensional, $\mathbf{z}$ la variable transformada, cuando se ha llevado a cabo la extracción de características, e y la variable discreta correspondiente a los datos.

Un importante resultado de la teoría de la información es el proporcionado por la desigualdad del tratamiento de datos [Cover, 2006], que nos dice que la información que un vector $\mathbf{x}$ contiene acerca de otra variable y no aumentará si se realiza una transformación determinista sobre $\mathbf{x}$, es decir

$$I(\mathbf{f}(\mathbf{x}), y) \leq I(\mathbf{x}, y) \tag{1.23}$$

Para el caso particular en que $\mathbf{f}(\mathbf{x})$ es una transformación sin pérdidas, se cumple la igualdad (por ejemplo, cuando la transformación es lineal, de la forma $\mathbf{f}(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$, y además $\mathbf{W}$ es de rango completo). Ya que no podemos aumentar esta información mutua mediante una transformación determinista (no se puede reducir la incertidumbre), el objetivo es reducir la dimensión conservando la mayor cantidad posible de información.

El concepto de suficiencia estadística define la capacidad de una transformación $\mathbf{f}(\mathbf{x})$ para contener toda la información de $\mathbf{x}$ que es relevante acerca de y. Se dice que $\mathbf{f}(\mathbf{x})$ es estadístico suficiente con respecto a y si se cumple la relación markoviana $\mathbf{x} \rightarrow \mathbf{f}(\mathbf{x}) \rightarrow y$. En tal caso, se preserva la totalidad de la información relevante, y la desigualdad en (1.23) se convierte en una igualdad.

La información mutua entre una determinada característica x_k y la variable de referencia y puede ser una buena medida de relevancia, ya que $I(x_k, y)$ mide la cantidad de información que x_k contiene acerca de y, y viceversa. Hay resultados en el ámbito de la teoría de la información que relacionan probabilidad de error (o error de Bayes) e información mutua. Uno de estos resultados es la cota de Fano [Cover, 2006]

$$p_e \geq \frac{H(y) - I(\mathbf{x}, y) - 1}{\log(N_c)}$$

que establece una cota inferior al error de clasificación, a partir de la información mutua entre el conjunto de variables a clasificar $\mathbf{x}$ y la clase y, y de la entropía de la clase $H(y)$. Otro resultado importante es la cota superior de Hellman-Raviv [Hellman, 1970], que viene dada por

$$p_e \leq \frac{1}{2}\left(H(y) - I(\mathbf{x}, y)\right)$$

Ambas cotas tienen en común el hecho de que disminuyen con $I(\mathbf{x}, y)$, lo que parece justificar el objetivo intuitivo de buscar variables o proyecciones con una alta información.

El uso de la entropía y la información mutua de Renyi ha dado lugar a resultados similares [Erdogmus, 2004], con cotas que se reducen cuando esta información crece.

De acuerdo con estos resultados, es razonable el uso de la información mutua como criterio de selección o extracción de características si se quiere hacer mínimo el error de clasificación. En cualquier caso, estos resultados han de ser interpretados con la reserva oportuna. Se ha demostrado que no existe ningún estimador del error de Bayes que sea válido para cualquier distribución, incluso cuando el número de muestras tiende a infinito, mediante el siguiente teorema formulado en [Antos, 1999]:

Teorema 1.1 *Sea $\phi_n(\mathbf{X}, \mathbf{Y})$ un estimador del error de Bayes a partir de una secuencia $\mathbf{X}$ de n muestras de $\mathbf{x}$ y de sus correspondientes etiquetas $\mathbf{Y}$. Sea ϵ el error real y $\{a_n\}$ una secuencia de números positivos que converge a cero. Entonces, existe una distribución de $\mathbf{x}$ e y que cumple*

$$E\{|\phi_n(\mathbf{X}, \mathbf{Y}) - \epsilon|\} \geq a_n$$

Este resultado es de vital importancia, ya que demuestra la inexistencia de un método de estimación del error asintótica y universalmente mejor que el resto. Puede ser interpretado en términos de selección de características, ya que se establece que

ningún método de reducción de la dimensión podrá usar como referencia el error de Bayes para demostrar su superioridad universal respecto de los demás.

1.3.3. La estimación de la densidad de probabilidad

En los anteriores apartados ha podido constatarse que las herramientas y conceptos de la teoría de la información necesitan disponer de la FDP de las variables con las que se trabaja. En un ámbito de aprendizaje basado en muestras, los datos se interpretan como realizaciones de una variable aleatoria con una densidad desconocida. En aprendizaje, se asume una diferenciación entre los métodos generativos y los discriminativos. Los primeros hacen uso de modelos de la FDP de los datos, mientras que los últimos no necesitan dichos modelos. La estimación de la FDP de los datos es necesaria si se pretende hacer uso de la información mutua, o alguna otra de las divergencias explicadas, como criterio para reducir la dimensión.

Los métodos propuestos en la literatura para modelado de FDPs se suelen dividir en tres grupos: paramétricos, semi-paramétricos y no paramétricos. Los métodos paramétricos son los más sencillos de usar y asumen que la densidad desconocida pertenece a una determinada familia de FDPs con descripción analítica. La tarea de estimar la FDP se reduce, por tanto, a encontrar los parámetros del modelo que mejor se ajusten a los datos. Este ajuste suele llevarse a cabo mediante un criterio de máxima verosimilitud [Bishop, 1995].

Los modelos semiparamétricos no se restringen a una forma funcional en concreto, por lo que son más flexibles. Los denominados modelos de mezcla responden a esta descripción y consisten en mezclas de modelos paramétricos que tienen la forma

$$\hat{p}(\mathbf{x}) = \sum_{l=1}^{L} \alpha_l f(\mathbf{x}|\boldsymbol{\theta}_l) \tag{1.24}$$

donde cada α_l es la probabilidad a priori de que $\mathbf{x}$ pertenezca al grupo l. Se cumple que $\sum_l \alpha_l = 1$. Cada submodelo está caracterizado por un conjunto de parámetros $\boldsymbol{\theta}_l$.

En concreto, los modelos basados en mezclas de gausianas (*Gaussian Mixture Models*, GMM) constituyen la familia más extendida de modelos de mezcla. Un modelo GMM asume un agrupamiento en los datos, de forma que cada muestra pertenece a un grupo que responde a un modelo de tipo paramétrico gausiano. El problema de estimar simultáneamente los α_l y los $\boldsymbol{\theta}_l$ es resuelto satisfactoriamente por el algoritmo de promediado-maximización (*Expectation-Maximization*, EM) [Bishop, 1995]. En aprendizaje basado en teoría de la información, los GMM tienen el problema de que si cambiamos el conjunto de variables a utilizar, en un esquema secuencial de selección o extracción de características, se exige reajustar el modelo en cada iteración, que es algo que puede llegar a ser computacionalmente muy costoso. Además persiste, aunque en menor grado, la dependencia de la validez del modelo con la familia de modelos paramétricos considerados para la mezcla.

Estas limitaciones han hecho que los modelos no paramétricos sean más utilizados en aprendizaje basado en teoría de la información. El esquema no paramétrico más difundido es el modelo de Parzen [Fukunaga, 1990]. Este tipo de modelado recibe también el nombre de densidad basada en núcleos (*kernel density estimation*, KDE). Un modelo de Parzen tiene la forma

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} k(\mathbf{x} - \mathbf{x}_i) \tag{1.25}$$

donde $k(\cdot)$ es la ventana o *kernel* utilizada. La función ventana tiene las mismas propiedades que una FDP, es decir, es no negativa y tiene una integral unitaria a lo largo de su dominio. El modelo se construye a partir de un conjunto de las muestras o centroides $\mathbf{x}_i$, que son realizaciones de la variable aleatoria cuya FDP se pretende modelar. En estos puntos es donde están centradas las ventanas.

Lo que hace que este modelo sea tan ampliamente usado es el hecho de que dependa únicamente de las muestras, y de la elección de la función núcleo. En realidad, a pasar de su denominación de método no paramétrico, en un modelo de Parzen hay que precisar el tipo de ventana usado así como su ancho. Un inconveniente de estos modelos es el alto coste de evaluar la probabilidad en un punto $\mathbf{x}$ determinado, ya

que exige la evaluación de N ventanas, lo que puede ser intensivo para bases de datos con tamaños muestrales grandes.

1.3.4. Aprendizaje basado en teoría de la información

A continuación se realiza una revisión de los métodos más relevantes que han sido propuestos en la literatura para aprendizaje basado en teoría de la información, tanto de forma no supervisada como supervisada.

Aprendizaje no supervisado

Los conceptos de la teoría de la información fueron aplicados a la construcción de modelos estadísticos, usando la entropía como criterio para penalizar el número de parámetros libres en el modelo. Ejemplos de estos criterios son el criterio de información de Akaike (AIC) y el criterio de información Bayesiano (*Bayesian Information Criterion*, BIC) [Burnham, 1998]. Las técnicas de aprendizaje más recientes han hecho que estos criterios se reformulen para servir como mecanismos de regularización de SVM [Kobayashi, 2006].

Los principios de la teoría de la información fueron aplicados por primera vez al aprendizaje de tipo neuronal desde una perspectiva no supervisada. En [Linsker, 1988] se proponen unas reglas de aprendizaje autoorganizado que hacen máxima la información mutua entre la entrada y la salida de un red neuronal. Más recientemente se ha propuesto un esquema para el entrenamiento de SOMs mediante un criterio de máxima información mutua en la fase de competición [Kamimura, 2006] en lugar del criterio clásico de proximidad euclídea. Por otro lado, las técnicas de búsqueda de la proyección óptima (*projection pursuit*) incorporaban el concepto de información y entropía al problema de encontrar las proyecciones más relevantes o informativas de un conjunto de datos hacia una dimensión menor [Huber, 1985]. Esta familia de técnicas constituyen el precedente directo de los métodos de análisis de componentes independientes (*Independent Component Analysis*, ICA). Para ICA se han descrito diferentes soluciones, y la mayor parte de ellas hace uso de contrastes basados

en teoría de la información, ya que el concepto de independencia estadística entre dos variables se describe en términos de su información mutua. Otros métodos ICA actúan sobre una función de verosimilitud sobre los datos, aunque puede demostrarse la equivalencia entre verosimilitud e información mutua [Hyvarinen, 2001].

El problema ICA parte de un conjunto de señales o variables unidimensionales $\mathbf{x}$ que se asume generado mediante la mezcla de una serie de señales estadísticamente independientes $\mathbf{s}$, de la forma $\mathbf{x} = \mathbf{A}^T\mathbf{s}$. El objetivo es, por tanto encontrar la matriz desconocida $\mathbf{A}$ y, mediante su inversión, un conjunto de señales $\mathbf{z}$ que sea igual a $\mathbf{s}$ salvo incertidumbre en el orden de las variables o su escalado. En términos de información mutua, el problema consiste en obtener la matriz $\mathbf{W}$ que invierta la transformación $\mathbf{A}$. El criterio es, por tanto

$$\mathbf{W}_{ICA} = \arg\min_{\mathbf{W}} I(\mathbf{z}) = \arg\min_{\mathbf{W}} I(\mathbf{W}^T\mathbf{x}) \tag{1.26}$$

donde $I(\mathbf{z})$ está definida como

$$I(\mathbf{z}) = I(z_1, z_2, \ldots, z_D) = \sum_{k=1}^{D} h(z_k) - h(\mathbf{z}) \tag{1.27}$$

La forma de abordar la minimización de esta magnitud difiere para los distintos métodos propuestos. El criterio InfoMax [Bell, 1995] propone maximizar el segundo término de (1.27). Para ello, se dispone una capa de unidades no lineales (sigmoidales) a la salida de la mezcla de separación, y la forma de maximizar la entropía es encontrar los pesos y centros que hagan que la FDP conjunta a la salida sea uniforme, que es la distribución acotada de mayor entropía. Los pesos constituyen los elementos de la matriz de separación.

Otros esquemas, como FastICA actúan sobre el primer miembro en (1.27): minimizar las entropías individuales garantiza que se minimice $I(\mathbf{z})$, ya que $h(\mathbf{z})$ es invariante cuando el determinante de $\mathbf{W}$ es unitario, dado que se cumple [Hyvarinen, 2001]

$$h(\mathbf{z}) = h(\mathbf{x}) + \log|\det(\mathbf{W})|$$

El uso de criterios de entropía e información mutua a la salida un perceptrón multicapa (*MultiLayer Perceptron*, MLP) ha sido propuesto en [Principe, 2000] para la reducción no lineal de la dimensión. La entrada y salida del MLP están relacionadas por una función no lineal $\mathbf{z} = g(\mathbf{w}^T\mathbf{x})$. El primero de los criterios propuestos es el de salida con máxima entropía. La entropía de salida se mide mediante su grado de parecido con una densidad uniforme. Más concretamente, mediante el error cuadrático integrado entre ambas

$$J = \frac{1}{2}\int_D (u(\mathbf{z}) - \hat{p}(\mathbf{z}))^2 d\mathbf{z} \tag{1.28}$$

La integral en (1.28) se puede sustituir por un muestreo en una malla de N_P puntos, de la forma

$$J \approx \sum_{j=1}^{N_P} \frac{1}{2}(u(\mathbf{z}_j) - \hat{p}(\mathbf{z}_j))^2 \Delta\mathbf{z}$$

donde la densidad $\hat{p}(\mathbf{z})$ está estimada mediante un modelo de Parzen. La optimización del coste J tiene lugar mediante un ascenso por gradiente. La derivada viene dada por

$$\frac{\partial J}{\partial \mathbf{w}} = \sum_{i=1}^{N} \frac{\partial J}{\partial \mathbf{z}_i}\frac{\partial \mathbf{z}_i}{\partial \mathbf{w}}$$

Las proyecciones no lineales obtenidas por el MLP pueden ser interpretadas como direcciones de máxima incertidumbre. Esto supone una generalización de PCA, ya que el criterio de máxima varianza es extendido a uno de máxima entropía.

En el mismo trabajo se propone otra alternativa para obtener máxima entropía a la salida, usando la misma arquitectura neuronal. Lo que se maximiza es la entropía de Renyi, tal y como se describió en (1.18). Para ello se sigue haciendo uso de modelos de Parzen para modelar las FDPs. También se propone un esquema para separación ciega de fuentes no lineal, basado en la divergencia de Cauchy-Schwartz (1.22).

La ventaja de trabajar con las divergencia de Renyi, cuadrática o de Cauchy-Schwartz es que las integrales de las FDPs cuadráticas tienen solución analítica aplicando la propiedad de convolución de gausianas, según la cual

$$\int G(\mathbf{x} - \mathbf{x}_i|\sigma^2)G(\mathbf{x} - \mathbf{x}_j|\sigma^2)d\mathbf{x} = G(\mathbf{x}_i - \mathbf{x}_j|2\sigma^2)$$

Entonces podemos hacer el desarrollo

$$\begin{aligned}\int \hat{p}^2(\mathbf{x})d(\mathbf{x}) &= \int \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{x}-\mathbf{x}_i|\sigma^2)\frac{1}{N}\sum_{j=1}^{N} G(\mathbf{x}-\mathbf{x}_j|\sigma^2)d\mathbf{x} \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int G(\mathbf{x}-\mathbf{x}_i|\sigma^2)G(\mathbf{x}-\mathbf{x}_j|\sigma^2)d\mathbf{x} \qquad (1.29) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\mathbf{x}_i-\mathbf{x}_j|2\sigma^2)\end{aligned}$$

La información mutua ha sido usada satisfactoriamente en [Yang, 2001] para paliar la "maldición de la dimensión" en la construcción de modelos de mezcla de gausianas. Para ello, se realiza una purga de variables. El algoritmo decremental elimina progresivamente las variables que tienen una alta dependencia de las demás componentes.

Aprendizaje supervisado

Algunos de los esquemas de selección y extracción de características propuestos en la literatura realizan una revisión del planteamiento básico ICA para adaptarlo a un entorno supervisado. El método ICA condicional (Conditional ICA, CICA) [Akaho, 2002] formula ICA a partir de las probabilidades condicionadas. La información mutua a maximizar está condicionada a y, es decir

$$I(\mathbf{z}|y) = D_{KL}\left(p(\mathbf{z}|y), \prod_{k=1}^{D} p(z_k|y)\right)$$

Para la estimación de las probabilidades condicionadas se hace uso de modelos de Parzen.

Otra versión supervisada de ICA la encontramos en [Kwak, 2002b], donde se propone un modelo de mezclas alternativo. Se asume un modelo según el cual los datos han sido generados mediante una mezcla de dos conjuntos de señales: uno de ellos está relacionado con las clases (es estadísticamente relevante) y el otro no (es irrelevante). El proceso de separación toma como entrada el vector compuesto por $\mathbf{x}$ y por la clase, y genera a la salida un vector $\mathbf{z}_r$ de características relevantes y

otro $\mathbf{z}_{ir}$ de características irrelevantes. Se establece la restricción sobre la matriz de separación $\mathbf{W}$ según la cual los elementos que relacionan y con el vector de salida $\mathbf{z}_{ir}$ son nulos.

El principio de máxima relevancia y mínima redundancia, cuando opera sobre datos discretos para selección o extracción de características, puede ser planteado en términos de codificación. El método del "cuello de botella" de información (*Information Bottleneck*, IB) busca una asignación entre cada elemento $\mathbf{x}$ y su palabra-código $\tilde{\mathbf{x}}$ de forma que se lleve a cabo una reducción eficiente de los datos [Tishby, 1999]. Al plantearse en un contexto de aprendizaje supervisado, se persigue que $\tilde{\mathbf{x}}$ transporte el máximo de información posible acerca de la variable auxiliar y. Por tanto, IB puede ser interpretado como un método de agrupamiento supervisado. El funcional a minimizar es

$$\hat{f}(\tilde{\mathbf{x}}|\mathbf{x}) = \arg \min_{f(\tilde{\mathbf{x}}|\mathbf{x})} [I(\tilde{\mathbf{x}}, \mathbf{x}) - \beta I(\tilde{\mathbf{x}}, y)]$$

donde se explicita la búsqueda de máxima relevancia y mínima redundancia. El principio de IB surge como una revisión supervisada de la teoría de la tasa de distorsión (*Rate Distortion Theory*) de Shannon y Kolmogorov, que busca la codificación $f(\tilde{\mathbf{x}}|\mathbf{x})$ que maximiza la relevancia mientras minimiza la distorsión introducida por la asignación [Cover, 2006].

El mismo principio de máxima relevancia y mínima redundancia se aplica en [Fleuret, 2004] al problema de la selección de características, planteando el concepto de cubierta de Markov en un contexto de teoría de la información. El algoritmo, incremental, va construyendo un conjunto S de variables mediante el procedimiento

$S(1) = \arg \max_j I(y, x_j)$
Para $k = 2$ hasta d
    $S(k) = \arg \max_j \left[\min_{l \leq k} I(y, x_j | x_{S(l)})\right]$
Fin

Como puede verse, el algoritmo tiene en cuenta tanto la relevancia como la redundancia, ya que el valor de $I(y, x_i|x_{S(l)})$ es bajo si x_j no proporciona información

sobre y o bien esa información ha sido ya captada por alguna de las variables ya seleccionadas $x_{S(l)}$. Dado que el algoritmo está propuesto para problemas de clasificación con datos binarios, la estimación de la información mutua es sencilla ya que opera sobre entidades discretas.

Los esquemas de selección de características han incorporado la información mutua como criterio en numerosos trabajos dentro de la literatura. En [Battiti, 1994] se planteaba un esquema incremental según el cual al conjunto de variables seleccionadas S se incorpora en cada iteración la variable que maximiza la información mutua dada por $I(x_k, y)$, pero que hace mínima $I(x_k, S)$. Es decir, se trata de alcanzar un equilibrio entre relevancia y redundancia tal y como se describió en el Apartado 1.2. La variable x_k a incorporar al conjunto S es la que hace máximo el criterio

$$I(x_k, y) - \beta I(x_k, S)$$

Para estimar las $I(x_k, y)$ y $I(x_k, S)$, se recurre a la discretización de cada variable, para poder recurrir al sumatorio en (1.3.1) en lugar de la integral en (1.16).

El uso de un estimador de Parzen junto con un estimador muestral de la entropía ha sido propuesto en [Kwak, 2002a]. La información mutua a maximizar viene expresada de la forma

$$I(S, y) = H(y) - H(y|S) \tag{1.30}$$

Dado que $H(y)$ es un término constante, el objetivo de maximizar la información mutua en (1.30) es equivalente a tratar de minimizar la entropía $H(y|S)$. Esta última es estimada de la forma

$$\hat{H}(y|\mathbf{X}) = -\sum_{i=1}^{N} \frac{1}{N} \sum_{l=1}^{N_c} \hat{p}(c_l|\mathbf{x}_i) \log \hat{p}(c_l|\mathbf{x}_i)$$

Para obtener la información mutua, se ha sustituido la integral a lo largo de todo el dominio de $\hat{p}(y|\mathbf{x})$ por un sumatorio en los puntos determinados por los $\mathbf{x}_i$. Cada estimador no paramétrico $\hat{p}(c_l|\mathbf{x})$ tiene la forma

$$\hat{p}(c_l|\mathbf{x}) = \frac{\sum_{i \in I_l} G(\mathbf{x} - \mathbf{x}_i|\sigma^2)}{\sum_{l'=1}^{N_c} \sum_{j \in I_{l'}} G(\mathbf{x} - \mathbf{x}_j|\sigma^2)} \tag{1.31}$$

siendo I_l el conjunto de índices de las muestras pertenecientes a la clase c_l. Al igual que en el algoritmo propuesto en [Battiti, 1994], la búsqueda de características es incremental.

El mismo esquema secuencial e incremental ha sido también propuesto en [Chow, 2005]. Los autores hacen uso de nuevo del estimador de Parzen, pero en lugar de un estimador muestreado de la entropía, se recurre a la divergencia de Cauchy-Schwartz que se vio en (1.22). La ventaja del uso de esta divergencia es que la integral es fácilmente resoluble para modelos no paramétricos, como se vio en (1.29). Siguiendo este procedimiento se calcula cada término involucrado en la estimación de $I_{CS}(S, y)$ que queda de la forma

$$\begin{aligned} I_{CS}(\mathbf{x}_S, y) = &\log \sum_{l=1}^{N_c} \int_{\mathbf{x}_S} p^2(\mathbf{x}_S, c_l) d\mathbf{x}_S + \log \sum_{l=1}^{N_c} P^2(c_l) \log \int_{\mathbf{x}_S} p^2(\mathbf{x}_S) d\mathbf{x}_S \\ &- 2 \log \sum_{l=1}^{N_c} \sum_{j=1}^{N_c} \int p(\mathbf{x}_S, c_l) P(c_j) p(\mathbf{x}_S) d\mathbf{x}_S \end{aligned} \tag{1.32}$$

El principio general de maximizar la relevancia de las variables obtenidas a la par que se minimiza su redundancia ha sido también propuesto en términos de información mutua explícitamente en [Peng, 2005]. Los autores proponen la maximización del contraste

$$\max_S \left[D(S, y) - R(S) \right] \tag{1.33}$$

donde $D(S, y)$ es la función de dependencia de las variables seleccionadas respecto de las clases, y $R(S)$ es la medida de redundancia entre las variables. Estas magnitudes vienen dadas por

$$\begin{aligned} D &= \frac{1}{|S|} \sum_{f_k \in S} I(x_{f_k}, y) \\ R &= \frac{1}{|S|^2} \sum_{f_k, f_j \in S} I(x_{f_k}, x_{f_j}) \end{aligned} \tag{1.34}$$

Los autores demuestran que la aplicación de una búsqueda secuencial con este contraste conduce a la misma solución que si se maximizara la información mutua

dada por $I(S, y)$. El inconveniente de trabajar con $I(S, y)$ es que implica el uso de un modelo de densidad de probabilidad multidimensional, mientras que el criterio expresado en (1.33) y (1.34) requiere únicamente el uso de estimaciones unidimensionales.

En el Apartado 1.3.4 se recogen los esquemas propuestos en [Principe, 2000] para reducción de la dimensión no supervisada con un MLP y criterios basados en teoría de la información. En el mismo trabajo se describe un sistema de aprendizaje supervisado con MLP, en el que, a diferencia del esquema ICA, que trata de minimizar la información mutua entre las componentes de salida, ahora se maximiza respecto a una variable de referencia. Para la estimación de la información mutua se recurre de nuevo a la divergencia de Cauchy-Schwartz.

En [Torkkola, 2003] se propone un método para extracción lineal supervisada que hace uso de la divergencia cuadrática en (1.20). La pseudo-información mutua a calcular equivale a la divergencia $D_Q(p(\mathbf{x}, y), p(\mathbf{x})P(y))$, lo que, según (1.22), da lugar a

$$I_Q(\mathbf{x}, y) = \sum_{l=1}^{N_c} \int_{\mathbf{x}} p^2(\mathbf{x}, c_l) d\mathbf{x} + \sum_{l} \int_{\mathbf{x}} P^2(c_l) p^2(\mathbf{x}) d\mathbf{x} - 2 \sum_{l} \int_{\mathbf{x}} p(\mathbf{x}, y) P(c_l) p(\mathbf{x}) d\mathbf{x} \tag{1.35}$$

Las integrales en (1.35) son resolubles analíticamente dado que implica la convolución de gausianas (la estimación de las densidades $p(\mathbf{x}, y)$ y $p(\mathbf{x})$ son modelos de Parzen, mientras que la $P(y)$ es una distribución empírica), como se vio en (1.29). La interacción de cada par de muestras recibe el nombre de "potencial", por su analogía física. La derivada de cada potencial viene dada por

$$\frac{\partial}{\partial \mathbf{x}_i} G(\mathbf{x}_i - \mathbf{x}_j | 2\sigma^2) = \frac{\mathbf{x}_j - \mathbf{x}_i}{2\sigma^2} G(\mathbf{x}_i - \mathbf{x}_j | 2\sigma^2)$$

mientras que la derivada de la $I_Q(\mathbf{x}, y)$ respecto a la matriz de transformación es combinación lineal de las aportaciones de cada muestra

$$\frac{\partial I_Q}{\partial \mathbf{W}} = \sum_i \frac{\partial I_Q}{\partial \mathbf{z}_i} \frac{\partial \mathbf{z}_i}{\partial \mathbf{W}}$$

Para restringir la búsqueda de soluciones y reducir los grados de libertad del problema, los autores proponen aplicar condiciones de ortogonalidad y normalización a los vectores de la matriz $\mathbf{W}$. Para imponer ortonormalidad en la matriz $\mathbf{W}$, se propone reparametrizar mediante rotaciones de Givens, es decir, expresar $\mathbf{W}$ mediante un conjunto de rotaciones angulares de la forma

$$\mathbf{W} = \mathbf{W}_s \left(\prod_{k=1}^{d} \prod_{j=d+1}^{D} \mathbf{R}(k, j, \theta_{kj}) \right)$$

donde cada $\mathbf{R}(k, j, \theta_{kj})$ es una matriz que produce un giro de θ_{kj} radianes entre las componentes k y j. Los pares de variables para cada abatimiento consta de una componente que está entre las d seleccionadas a la salida y otra de las $D - d$ que no lo están. Al ser cada matriz de rotación ortonormal, el producto de todas ellas también lo es. La matriz $\mathbf{W}_s$ lleva a cabo la selección de las d proyecciones de salida (consta de las M primeras filas de una matriz identidad $D \times d$).

La reparametrización entre los elementos de $\mathbf{W}$ y los ángulos θ_{kj} se realiza de la forma

$$\frac{\partial I_Q}{\partial \theta_{kj}} = \sum_{n,m} \frac{\partial I_Q}{\partial w_{nm}} \frac{\partial w_{nm}}{\partial \theta_{kj}}$$

Esta reparametrización se ha usado también en [Hild, 2006], junto a un estimador de la información mutua como el de (1.16), y un estimador adaptativo de la entropía, para la extracción de características.

Los modelos no paramétricos también han servido para construir un esquema de máxima verosimilitud cuya equivalencia con la información mutua está asintóticamente probada [Peltonen, 2005]. Se utiliza la siguiente verosimilitud

$$L = \sum_{i=1}^{N} \log \hat{p}(c_{y_i} | \mathbf{f}(\mathbf{x}_i))$$

donde la transformación es lineal, $\mathbf{f}(\mathbf{x}) = \mathbf{W}^T \mathbf{x}$, y el modelo $\hat{p}(c|\mathbf{f}(\mathbf{x}))$ es idéntico al definido en (1.31). La maximización de esta verosimilitud lleva a la construcción de la transformación $\mathbf{W}$ que permite una mayor discriminación entre clases, con la ventaja de que los modelos construidos $\hat{p}(c_l|\mathbf{W}^T\mathbf{x})$ permiten realizar aprendizaje de tipo generativo.

1.4. Objetivos y Contenido de la tesis

La presente tesis aborda el problema de la extracción lineal de características mediante criterios basados en teoría de la información. Para ello, se plantean técnicas que tratan de superar algunos de los problemas que se presentan en la estimación de magnitudes propias de esta teoría. Una de las principales limitaciones de los métodos presentes en la literatura es su alta dependencia respecto a una precisa estimación de la función densidad de probabilidad. Cuando esta densidad se utiliza para evaluar el valor de una entropía, una verosimilitud o una información mutua, se incurre en una doble imprecisión: la correspondiente al modelado de la función de densidad y la cometida al estimar el valor de la magnitud deseada a partir de ésta. Para aliviar esta problemática, se proponen dos líneas de trabajo claramente diferenciadas.

En la primera línea de trabajo, se elude el problema de la estimación de la densidad de probabilidad, recurriéndose a las técnicas basadas en expansiones mediante momentos. La estadística clásica cuenta con estimaciones polinomiales que modelan el valor de la densidad a partir de la aproximación de su grado de disparidad con la densidad gausiana. Al ser estas expansiones polinomiales, el valor de la estimación es fuertemente dependiente de los valores más destacados de la variable, aquellos procedentes de las colas de la distribución. En el Capítulo 2 se plantea una estimación de la entropía basada en expansiones no polinomiales y se propone un método para la maximización de la información mutua entre las proyecciones obtenidas y las etiquetas. Se incorpora una discusión teórica sobre el valor de la imprecisión cometida en la estimación de la información mutua, y se lleva a cabo un conjunto de experimentos que ponen de relieve la capacidad del método para extraer con éxito características relevantes en problemas donde la función de discriminación presenta fuertes no linealidades.

En la segunda línea de investigación propuesta en la tesis, se aborda el problema del modelado de la FDP de los datos mediante modelos de Parzen. En el Capítulo 3 se proponen algoritmos para el modelado de Parzen con ventana gausiana, con dis-

tintos grados de complejidad de la matriz de covarianza. Los algoritmos propuestos están basados en la maximización de la verosimilitud de los mismos, lo que requiere hacer uso de un esquema de validación cruzada. Al ser la verosimilitud una magnitud estrechamente vinculada con la teoría de la información, nos permitirá relacionar el diseño de los modelos con posteriores resultados de su aplicación. En concreto, se incorpora un resultado teórico que demuestra la optimalidad de los modelos obtenidos por dichos algoritmos para la estimación de la entropía. En el Capítulo 4 se proponen algunas aplicaciones de los modelos de Parzen, optimizados de acuerdo con el método propuesto en el Capítulo 3, a distintos problemas de aprendizaje. Se propone un método de extracción de características que hace uso de estos modelos junto a conceptos de teoría de la decisión, así como otro método que hace uso de modelos GMM para la maximización de la información mutua, también con el objetivo de la extracción de características relevantes. Junto a estas técnicas, también se plantean algoritmos para análisis de componentes independientes así como clasificación de Parzen.

En el Capítulo 5 se recogen las principales conclusiones que pueden extraerse del trabajo expuesto, y se proponen algunas líneas de trabajo que pueden considerarse la continuación natural de los métodos presentados a lo largo de la tesis.

Capítulo 2

Extracción de características basada en funciones de contraste

En el Apartado 1.3 se ha expuesto una revisión de los métodos más relevantes que han tratado de aplicar los conceptos de teoría de la información a la extracción de características. Como ha podido apreciarse, un rasgo mayoritario en estos métodos es el modelado de la función densidad de probabilidad de los datos para, a partir de ella, estimar magnitudes tales como entropía o información mutua.

Los métodos ICA más populares en la literatura, como FastICA, InfoMax o JADE eluden el problema de la estimación de la FDP de los datos tratando de estimar la información mutua entre las variables de salida, $I(\mathbf{z})$, directamente a partir de los datos. En el caso de FastICA, la entropía asociada a cada señal es estimada a partir de momentos no polinomiales sobre los datos [Hyvarinen, 1998]. En el caso de JADE, estos momentos son polinomiales de orden superior [Cardoso, 1998]. InfoMax opera mediante reglas de aprendizaje anti-hebbianas deducidas a partir de una red neuronal cuya entropía a la salida se pretende maximizar [Bell, 1995].

Basándonos en estos procedimientos que evitan estimar la densidad de probabilidad, en el presente capítulo se desarrolla un método de extracción de características que usa la información mutua respecto a las clases como criterio a maximizar. Re-

ducido el problema a uno de estimación de entropía, se podrá extrapolar el trabajo llevado a cabo en ICA, y más concretamente en la metodología usada en FastICA, a este ámbito de aprendizaje supervisado, obviándose la estimación de la densidad de probabilidad de los datos.

En el siguiente Apartado se mostrará un modelo teórico de mezcla de señales sobre el que se apoya el método propuesto para extracción de características. En el Apartado 2.2 se describe la función de contraste propuesta, así como la metodología para su optimización. En el Apartado 2.3 se presenta el algoritmo propuesto para la extracción de características, que llamaremos MMI (*Maximization of Mutual Information*) tal y como aparece en [Leiva-Murillo, 2007a]. La evaluación del método, en comparación con otros métodos lineales, se presenta en el Apartado 2.4. En el Apartado 2.5 se propone una discusión de los resultados así como una serie de conclusiones sobre el método propuesto.

2.1. Modelo de mezcla

El esquema de extracción de características lineal que se propone parte de la suposición de que los datos han sido generados mediante el modelo de mezcla que se muestra en la Figura 2.1. Se considera que las componentes de los datos observados, dadas por cada x_k, han sido generadas mediante la mezcla lineal $\mathbf{x} = \mathbf{A}^T\mathbf{s}$. El vector de señales $\mathbf{s}$ tiene la forma $\mathbf{s} = \begin{pmatrix} \mathbf{s}^r \\ \mathbf{s}^n \end{pmatrix}$, de manera que aparecen desacopladas las señales relevantes que contienen información sobre y ($\mathbf{s}^r$), y que conforman el estadístico suficiente respecto de la clase, de las que no contienen información sobre la clase ($\mathbf{s}^n$). Al igual que en el modelo asumido para ICA, explicado en el Apartado 1.3, en el que los datos se suponen generados mediante mezcla de variables estadísticamente independientes, esta suposición supone más un artificio matemático que una realidad física. En numerosas aplicaciones de ICA, no se busca invertir una mezcla sino la simple obtención de un conjunto de variables estadísticamente independien-

tes que permita una apropiada interpretación de los datos o una reducción de la redundancia presente en los mismos.

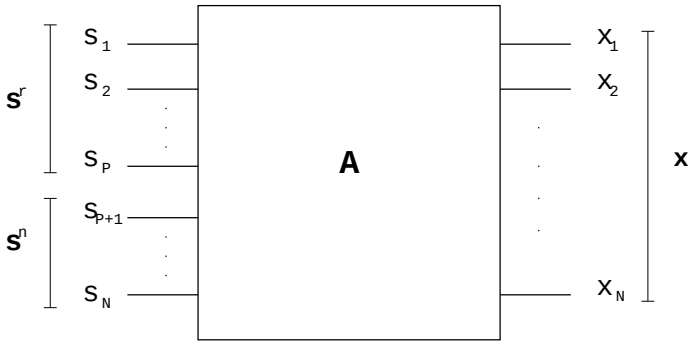

Figura 2.1: Modelo de mezcla. Se asume que $\mathbf{x}$ ha sido generada mediante una mezcla de variables relevantes e irrelevantes.

Para justificar la aplicación de este modelo al problema de la extracción de características basada en relevancia, necesitamos comprobar la siguiente igualdad

$$I(\mathbf{s}^r, y) = I(\mathbf{x}, y)$$

con la condición de que se cumpla $I(\mathbf{s}^n, y) = 0$. Si, como se ha dicho, $\mathbf{A}$ es invertible, se cumple que

$$I(\mathbf{x}, y) = I(\mathbf{s}, y) = I(\{\mathbf{s}^r, \mathbf{s}^n\}, y) = E_{\mathbf{s},y}\left\{\log \frac{p(\{\mathbf{s}^r, \mathbf{s}^n\}, y)}{p(\mathbf{s}^r, \mathbf{s}^n)P(y)}\right\}$$

que mediante la regla de Bayes se puede reexpresar como

$$I(\mathbf{x}, y) = E_{\mathbf{s}^r,\mathbf{s}^n,y}\left\{\log \frac{p(\mathbf{s}^n|\mathbf{s}^r, y)p(\mathbf{s}^r, y)}{p(\mathbf{s}^r, \mathbf{s}^n)P(y)}\right\} \tag{2.1}$$

De acuerdo con el teorema de suficiencia estadística [Cover, 2006], $\mathbf{s}^n$ es considerada irrelevante respecto de y si

$$p(\mathbf{s}^n|\mathbf{s}^r, y) = p(\mathbf{s}^n|\mathbf{s}^r)$$

En tal caso, se puede reescribir (2.1) como

$$I(\mathbf{x}, y) = E_{\mathbf{s},y}\left\{\log \frac{p(\mathbf{s}^n|\mathbf{s}^r)p(\mathbf{s}^r, y)}{p(\mathbf{s}^r, \mathbf{s}^n)P(y)}\right\}$$
$$= E_{\mathbf{s}^r,\mathbf{s}^n,y}\left\{\log \frac{p(\mathbf{s}^r, y)}{p(\mathbf{s}^r)P(y)}\right\}$$

La demostración finaliza mediante una marginalización con respecto a $\mathbf{s}^n$

$$I(\mathbf{x}, y) = E_{\mathbf{s}^r,y}\left\{\log \frac{p(\mathbf{s}^r, y)}{p(\mathbf{s}^r)P(y)}\right\} = I(\mathbf{s}^r, y)$$

El objetivo consiste, por tanto, en la obtención de la matriz de extracción $\mathbf{W}$ que permita recuperar la información relevante mediante la transformación $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, donde se espera que $\mathbf{z}$ sea equivalente a $\mathbf{s}^r$ con la única incertidumbre del orden de las señales, su signo o su escalado.

Por razones numéricas, conviene establecer condiciones sobre $\mathbf{W}$ que restrinjan el espacio de búsqueda. De forma similar al procedimiento seguido en ICA, forzaremos que $\mathbf{W}$ sea ortonormal, de forma que $\mathbf{W}^T\mathbf{W} = \mathbf{I}_D$. De esta manera, el ámbito de la búsqueda se limita al grupo algebraico de las matrices ortonormales en lugar del anillo de las matrices reales.

Antes de aplicar la restricción de la ortonormalidad, se necesita demostrar que una matriz ortonormal es suficiente para deshacer la mezcla y recuperar la separación entre características relevantes e irrelevantes. A continuación se demuestra que la condición no limita la posibilidad de alcanzar la separación entre $\mathbf{s}^r$ y $\mathbf{s}^n$.

Si la matriz de mezcla $\mathbf{A}$ es invertible, $\mathbf{A}^{-1}$ consigue la separación entre $\mathbf{s}^r$ y $\mathbf{s}^n$. Sea $\mathbf{A}^{-1} = \mathbf{QR}$ la descomposición QR de dicha matriz, de forma que $\mathbf{Q}$ es ortonormal y $\mathbf{R}$ es matriz diagonal inferior. In tal caso, $\mathbf{s} = \mathbf{A}^{-T}\mathbf{x} = \mathbf{R}^T\mathbf{Q}^T\mathbf{x}$. Sea $\mathbf{x}'$ definida como: $\mathbf{x}' = \mathbf{Q}^T\mathbf{x}$, de forma que $\mathbf{s} = \mathbf{R}^T\mathbf{x}'$. Dado que $\mathbf{R}^T$ es diagonal superior, el hecho de que en $\mathbf{s}$ estén perfectamente separadas las componentes relevantes e irrelevantes implica que también lo estarán en $\mathbf{x}'$. Eso significa que $\mathbf{Q}$ tendría la estructura $\mathbf{Q} = [\mathbf{W} \quad \mathbf{N}]$, siendo $\mathbf{W}$ la matriz que proyecta las características relevantes y $\mathbf{N}$ la que proyecta las irrelevantes.

En resumen, siempre existirá una matriz $\mathbf{W}$ de tamaño $N \times M$ y compuesta por vectores ortonormales que cumpla

$$I(\mathbf{x}, y) = I(\mathbf{W}^T\mathbf{x}, y) \tag{2.2}$$

a condición de que $\mathbf{x}$ haya sido generada conforme al modelo de mezcla de la Figura 2.1.

2.2. Función de coste propuesta

El problema de la estimación de $I(\mathbf{z}, y)$ es muy difícil debido a las razones enumeradas en el Apartado 1.3.1. En un problema de clasificación, en el que la variable auxiliar y es discreta, la expresión para la información mutua en (1.16) puede ser reescrita como

$$I(\mathbf{z}, y) = h(\mathbf{z}) - \sum_{l=1}^{N_c} P_y(c_l) h(\mathbf{z}|c_l)$$

donde c_l es cada una de las clases y $P_y(c_l)$ su probabilidad *a-priori*. A falta de información adicional, se podrán estimar estas probabilidades mediante la distribución empírica de las etiquetas de las muestras de entrenamiento. Debido a la dificultad de la estimación de cada una de las entropías involucradas, se propone la siguiente función de contraste alternativa

$$\max_{\{\mathbf{w}_k\}} \sum_{k=1}^{d} I(\mathbf{w}_k^T\mathbf{x}, y) = \max_{\{\mathbf{w}_k\}} \sum_{k} I(z_k, y) \tag{2.3}$$

Esto supone tomar como criterio la suma de las $I(z_k, y)$ individuales, dado que la estimación de la información mutua unidimensional es un problema más fácilmente abordable.

Ya que vamos a hacer uso del criterio en (2.3) en lugar del criterio ideal que sería el descrito en (2.2), es necesario obtener la relación entre ambas magnitudes. El sumatorio a optimizar en (2.3) puede ser expresado como

$$\sum_{k} I(z_k, y) = \sum_{k} \Big[h(z_k) - h(z_k|y) \Big] \tag{2.4}$$

Por otro lado, la información mutua correspondiente a la dependencia estadística entre las propias componentes de $\mathbf{z}$ es

$$I(\mathbf{z}) = \sum_k h(z_k) - h(\mathbf{z}) \tag{2.5}$$

y de forma similar, cuando está condicionada a y,

$$I(\mathbf{z}|y) = \sum_k h(z_k|y) - h(\mathbf{z}|y) \tag{2.6}$$

Combinando (2.4), (2.5) y (2.6) llegamos a

$$\sum_k I(z_k, y) = \Big[I(\mathbf{z}) + h(\mathbf{z})\Big] - \Big[I(\mathbf{z}|y) + h(\mathbf{z}|y)\Big]$$

Tras reagrupar términos, se puede reexpresar como

$$\sum_k I(z_k, y) = I(\mathbf{z}, y) + \Big[I(\mathbf{z}) - I(\mathbf{z}|y)\Big]$$

Hemos obtenido el error que cometemos al estimar la magnitud $I(\mathbf{z}, y)$ mediante la magnitud alternativa $\sum_k I(z_k, y)$. Su valor es $I(\mathbf{z}) - I(\mathbf{z}|y)$, y su cálculo o estimación no es trivial dado que supone tratar con la información mutua entre varias variables continuas. La estimación directa o indirecta de $I(\mathbf{z})$ ha sido ampliamente tratada en la bibliografía sobre métodos ICA, dado que en estos métodos representa la magnitud a minimizar. Una de las caracterizaciones que encontramos es la que hace uso de los coeficientes de correlación de orden superior llamados cumulantes [Cardoso, 1998].

En la práctica, los cumulantes de orden inferior o igual a cuatro son suficientes para caracterizar satisfactoriamente la información mutua en un esquema ICA. Dado que un cumulante de orden k implica tratar con un tensor de k dimensiones, el coste computacional que requiere su cálculo crece exponencialmente con el orden, y además la estimación se degrada por ser su valor muy sensible a muestras atípicas o, en general, sobresalientes [Hyvarinen, 2001].

Para reducir tanto $I(\mathbf{z})$ como $I(\mathbf{z}|y)$ se propone llevar a cabo un "blanqueado" o decorrelación mediante PCA sobre $\mathbf{x}$ de forma previa: de esta forma, los cumulantes

de primer y segundo orden serán cancelados para $\mathbf{x}$. Si, como aquí se propondrá, las proyecciones $\mathbf{w}_k$ están forzadas a tener norma unitaria y ser ortogonales entre sí, la decorrelación entre las variables de entrada será preservada a la salida, por lo que los cumulantes de orden uno y dos serán anulados también en $\mathbf{z}$.

2.3. El algoritmo de extracción de características MMI

A continuación mostramos el procedimiento para la estimación y la maximización de cada $I(z_k, y)$ en la función de coste en (2.3). La estimación tendrá lugar a partir de las entropías, como fue mostrado en (2.4). Por otro lado, cada entropía diferencial puede ser expresada de la forma

$$h(z) = h_G(z) - J(z) \tag{2.7}$$

donde $h_G(z)$ es la entropía si se asume gausianidad, es decir, la entropía que tendría z si fuese una gausiana con la misma varianza; $J(z)$ es la "negentropía" de z, y es un término acuñado en la literatura sobre ICA [Hyvarinen, 2001]. La negentropía mide el grado de disparidad entre la entropía de una variable aleatoria cualquiera y una gausiana. Es una magnitud no negativa, puesto que una variable gausiana es, de todas la que tienen igual varianza, la de mayor entropía. Algunos métodos ICA tratan de maximizar esta $J(z)$, bajo la suposición de que cuanto menos gausiana sea la señal proyectada, con mayor verosimilitud corresponderá a una fuente de información.

Las estimaciones de $J(z)$ más conocidas son las expansiones de Gram-Charlier y de Edgeworth [Haykin, 1998]. Estas expansiones se definen a partir de combinaciones lineales de momentos de orden superior sobre los datos.

2.3.1. La expansión polinomial de Gram Charlier

La expansión de Gram Charlier está basada en la función característica [Haykin, 1998], que se define como la transformada de Fourier de la FDP

$$\varphi_X(\omega) = \int_x p(x) e^{i\omega x} dx$$

Mediante la expansión de la exponencial mediante series de Fourier, se obtiene

$$\varphi_X(\omega) = 1 + \sum_{k=1}^{\infty} \frac{(i\omega)^k}{k!} m_k$$

donde cada m_k es el momento de orden k, que puede estimarse a partir de los datos, es decir

$$m_k = E\{x^k\} \approx \frac{1}{N} x_i^k$$

La propiedad más importante de $\varphi_X(\omega)$ es que proporciona una estimación $\hat{p}(x)$ de la densidad $p(x)$ mediante la transformada inversa de Fourier, de la forma

$$\begin{aligned} p(x) &= \mathcal{F}^{-1}\{\varphi_X(\omega)\} = \frac{1}{2\pi} \int \varphi_X(\omega) e^{i\omega x} d\omega \\ &= \alpha(x) \left[1 + \sum_k c_k H_k(x) \right] \end{aligned} \tag{2.8}$$

siendo $\alpha(x)$ una FDP gausiana normalizada, c_k los coeficientes dados por el agrupamiento de los momentos y $H_k(x)$ los polinomios de Hermite, que están definidos por la regla recursiva $H_1(x) = 1$ y $H_{k+1}(x) = xH_k(x) - kH_{k-1}(x)$. Para tratar con esta expansión de infinitos términos, se requiere su truncamiento en un determinado orden.

La entropía estimada de x viene dada por

$$h(x) = -\int_x p(x) \log p(x) dx$$

Dado que $\hat{p}(x)$ está definida a partir de una suma, la integral se hace irresoluble. Sin embargo, podemos hacer uso de la aproximación $\log(1+\epsilon) \approx \epsilon - \frac{\epsilon^2}{2}$ si suponemos

que el término dado por la suma en (2.8) es pequeño. De esta forma, se obtiene la siguiente estimación de la entropía

$$\begin{aligned}\hat{h}(x) \approx h_G(x) - \frac{\kappa_3^2}{12} - \frac{\kappa_4^2}{48} - \frac{(\kappa_6 + 10\kappa_3^2)^2}{1440} + \frac{3}{8}\kappa_3^2\kappa_4 + \frac{\kappa_3^2(\kappa_6 + 10\kappa_3^2)}{24} \\ + \frac{\kappa_4^2(\kappa_6 + 10\kappa_3^2)}{24} + \frac{\kappa_4(\kappa_6 + 10\kappa_3^2)^2}{64} + \frac{\kappa_4^3}{16} + \frac{(\kappa_6 + 10\kappa_3^2)^3}{432}\end{aligned} \tag{2.9}$$

siendo cada κ_k, por definición, el cumulante de orden k de x. Por ejemplo, $\kappa_3 = m_3$ y $\kappa_4 = m_4 - 3m_2^2$, denominándose κ_4 la kurtosis de x. La kurtosis es una magnitud importante en el ámbito de la estadística y la teoría de la probabilidad, ya que da una medida del grado de lo concentrada que está la FDP en torno a su media (es decir, de lo "picuda" que es la FDP).

La expresión en (2.9) da lugar, mediante el truncamiento en el cuarto orden, a la siguiente estimación de $J(\cdot)$

$$J(z) \approx \frac{1}{12}E\{z^3\}^2 + \frac{1}{48}\,\mathrm{kurt}(z)^2 \tag{2.10}$$

donde $\mathrm{kurt}(z)$ es la mencionada kurtosis.

La expansión de Edgeworth se obtiene de forma similar, y da lugar a una expansión de la función de densidad cuyos coeficientes decrecen con el orden de los momentos [Haykin, 1998]. Ambas expansiones tienen un número infinito de términos, por lo que el problema de su truncamiento (es decir, la elección del grado de la expansión, despreciándose los términos de orden mayor) es importante. Aun cuando el grado de truncamiento es bajo, como en el caso de (2.10), la estimación tiene una gran dependencia respecto a las muestras procedentes de las colas de la distribución, y es fuertemente sensible a muestras atípicas. Se han propuesto alternativas a Gram-Charlier para construir una estimación robusta de la densidad de probabilidad a partir de muestras que no presente estos problemas [Welling, 2005]. Sin embargo, no se ha resuelto el problema de la aplicación de este esquema a la estimación de la entropía.

2.3.2. Expansión alternativa y obtención de proyecciones individuales

Una alternativa eficiente y robusta a las expansiones polinomiales es la que se propone en [Hyvarinen, 2001], que consiste en el uso de combinaciones de funciones pares e impares del tipo

$$J(z) \approx k_1(E\{G_1(z)\})^2 + k_2(E\{G_2(z)\} - E\{G_2(\nu)\})^2 \tag{2.11}$$

donde $G_1(z)$ es una función impar y $G_2(z)$ una función par; ν es una variable aleatoria gausiana con la misma media y varianza que z. Por ejemplo, el uso de las funciones $G_1(z) = z^3$ y $G_2(z) = z^4$ dan lugar a la expresión polinómica en (2.10). Para paliar los mencionados inconvenientes de las expansiones polinómicas, se ha propuesto el uso de las funciones no polinomiales $G_1(z) = z\exp\left(\frac{-z^2}{2}\right)$ y $G_2(z) = \exp\left(\frac{-z^2}{2}\right)$ en aplicación a ICA, que se han mostrado robustas y precisas [Hyvarinen, 2001]. Esta elección da lugar a la siguiente estimación de $J(z)$

$$J(z) \approx k_1\left(E\{z\exp\left(\frac{-z^2}{2}\right)\}\right)^2 + k_2\left(E\{\exp\left(\frac{-z^2}{2}\right)\} - \sqrt{1/2}\right)^2 \tag{2.12}$$

con las constantes $k_1 = 36/(8\sqrt{3}-9)$ y $k_2 = 24/(16\sqrt{3}-27)$. Además de solventar los problemas de robustez mencionados, esta estimación exhibe un buen comportamiento numérico al ser optimizado, sin estancarse en mínimos locales.

Haciendo uso de (1.16) y (2.7), y teniendo en cuenta que $h(z_k|y) = \sum_l P_y(c_l)h(z_k|c_l)$, la información mutua entre cada proyección y las clases se define en términos de negentropía de la forma

$$I(z_k, y) = h_G(z_k) - J(z) - \sum_l P_y(c_l)\Big[h_G(z_k|c_l) - J(z_k|c_l)\Big] \tag{2.13}$$

La entropía de una gausiana viene determinada únicamente por la varianza, y su valor es

$$h_G(z) = \log((2\pi e)^{1/2}\sigma_z)$$

por lo que la expresión en (2.13) se puede desarrollar como

$$I(z_k, y) = \log((2\pi e)^{1/2}\sigma_{z_k}) - J(z) - \sum_l P_y(c_l)\left[\log((2\pi e)^{1/2}\sigma_{z_k|c_l}) - J(z_k|c_l)\right] \quad (2.14)$$

Si se fuerza que cada proyección esté normalizada, es decir $\|\mathbf{w}_k\| = 1$, se pueden llevar a cabo varias simplificaciones. Dado que la entrada ha sido blanqueada para eliminar los cumulantes de primer y segundo orden, quedando $\mathbf{C_x} = \mathbf{I}$ (donde $\mathbf{C_x}$ es la matriz de covarianzas de $\mathbf{x}$), tenemos que $\sigma_z^2 = \mathbf{w}^T\mathbf{C_x}\mathbf{w} = \mathbf{w}^T\mathbf{w} = 1$. Por tanto, la expresión (2.14) queda de la forma

$$I(z_k, y) = \sum_l P_y(c_l)J(z_k|c_l) - J(z_k) - \sum_l P_y(c_l)\log\sigma_{z_k|c_l}$$

La varianza de las componentes de salida vienen relacionadas con la varianza en la entrada mediante $\sigma_z^2 = \mathbf{w}^T\mathbf{C_x}\mathbf{w}$, donde $\mathbf{C_x}$ es la matriz de covarianzas de $\mathbf{x}$. La misma relación existe entre cada $\mathbf{C}_{\mathbf{x}|c_l}$ y cada $\sigma^2_{z|c_l}$.

Ahora estamos en disposición de calcular el gradiente, para llevar a cabo la optimización. Este vendría dado por

$$\nabla_{\mathbf{w}} I(z, y) = \sum_l P_y(c_l)\nabla_{\mathbf{w}} J(z|c_l) - \nabla_{\mathbf{w}} J(z) - \sum_l P_y(c_l)\frac{\mathbf{C}_{\mathbf{x}|c_l}\mathbf{w}}{\mathbf{w}^T\mathbf{C}_{\mathbf{x}|c_l}\mathbf{w}} \quad (2.15)$$

donde

$$\begin{aligned}\nabla_{\mathbf{w}} J(z) =& 2k_1 E\{z\exp(\frac{-z^2}{2})\}E\{\mathbf{x}(1-z^2)\exp(\frac{-z^2}{2})\} \\ & -2k_2\left(E\{\exp(\frac{-z^2}{2})\} - \sqrt{\frac{1}{2}}\right)E\{\mathbf{x}z\exp\frac{-z^2}{2}\}\end{aligned}$$

De forma equivalente, $\nabla_{\mathbf{w}} J(z|c_l)$ es obtenido a partir del subconjunto de muestras de $\mathbf{x}$ que pertenecen a la clase c_l. Llevaremos a cabo una optimización mediante un método de ascenso por gradiente estándar.

Ha de tenerse en cuenta que la diferencia fundamental entre el objetivo que se plantea aquí e ICA reside en el hecho de que ICA necesita un contraste cuyo valor sea cercano a cero en caso de estadística cercana a la gausiana, mientras que lo que se necesita para la extracción relevante de características es un contraste cuyo valor

sea creciente con el orden de dependencia estadística entre la proyección extraída y la clase, lo que supone un problema de mayor dificultad.

Como puede verse, la aplicación previa de PCA tal y como se propuso en el Apartado 2.2 lleva a cabo un control en la varianza de las componentes de salida, además de forzar que los vectores de proyección estén normalizados. El escalado de $J(z_k|c_l)$ con la varianza de su argumento puede sesgar la estimación de $I(z_k, y)$ de no estar dicha varianza normalizada.

Para visualizar la necesidad de esta normalización, en la Figura 2.2 se muestra el valor de la entropía estimada, así como de la real, para datos con estadística gausiana para un rango de valores de desviación típica entre 0.1 y 10 con 1000 muestras. Como puede apreciarse, la estimación es muy precisa cuando los datos están normalizados, pero se aleja bruscamente del valor real cuando la varianza deja de ser unitaria. La razón de esto es que los momentos exponenciales con los que se estima $J(z)$ no escalan de forma apropiada. Al no tener sentido la estimación propuesta para datos no normalizados, se tendrá en cuenta la propiedad del escalado de la entropía

$$h(az) = h(z) + \log a$$

De esta forma, se podrá aplicar la estimación de la siguiente forma

$$h(z) = h(\bar{z}) + \log \sigma_z = h_G(\bar{z}) + \log \sigma_z - J(\bar{z})$$

donde $\bar{z} = \dfrac{z}{\sigma_z}$.

La media de los datos es aun más influyente en la estimación (por culpa, de nuevo, del término $J(z)$) que la varianza. Esto es lógico, debido a la alta sensibilidad de los momentos exponenciales en (2.12) respecto a la media. En la Figura 2.3 se ilustra el problema para datos normalizados con varianza unitaria y estadística gausiana (para otros tipos de distribución se comporta de idéntica manera). La entropía real es constante ya que, como es sabido, es invariante a la media. Se exige por tanto tomar los momentos centrados sobre los datos, es decir, una vez se ha eliminado la media a los mismos.

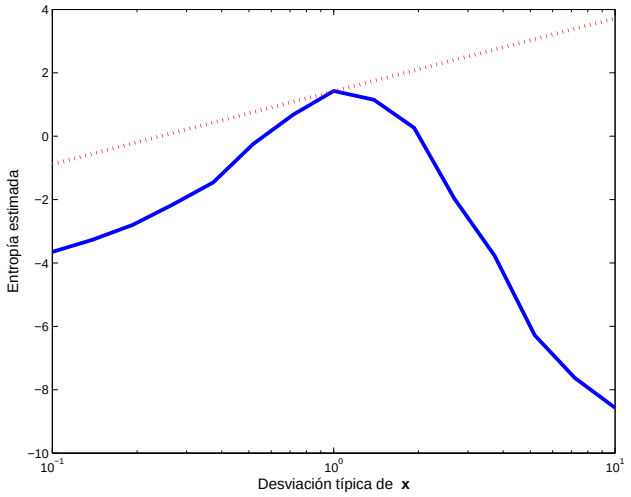

Figura 2.2: Entropía estimada (línea continua) y real (punteada) para datos unidimensionales con estadística gausiana en función de la desviación típica de los datos.

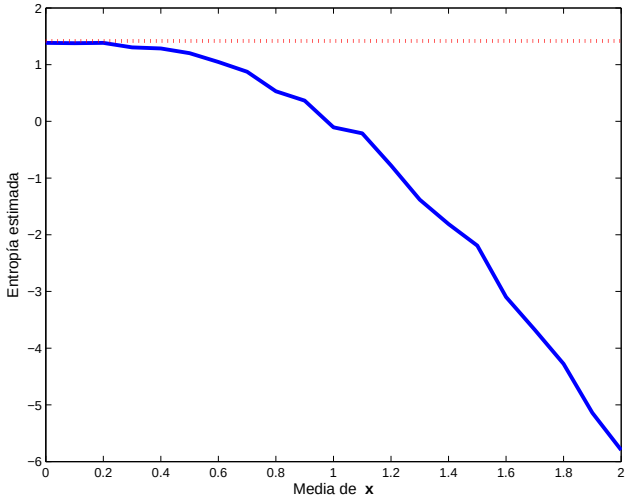

Figura 2.3: Entropía estimada (línea continua) y real (punteada) para datos unidimensionales con estadística gausiana en función de la media de los datos.

En cuanto al valor de la información mutua, se muestra la estimación en la Figura 2.4 para datos generados a partir de dos gausianas, siendo y la variable que indica la pertenencia de z a una u otra gausiana. Se ha realizado un barrido en el grado de separación entre las muestras correspondientes a una y otra clase. El valor real de la información mutua tiende a $\log 2$ en todos los casos: este valor se corresponde con la entropía de y ya que al estar las clases no solapadas no tenemos incertidumbre sobre y conocido z. El hecho de que el valor estimado siga creciendo se debe a la imprecisión que introduce $J(\cdot)$ cuando se aplica a datos multimodales. El estimador de $h(z)$ está sesgado a valores altos, lo cual puede ser positivo porque se tenderá a acentuar la multimodalidad de z y la unimodalidad de cada $z|c_l$.

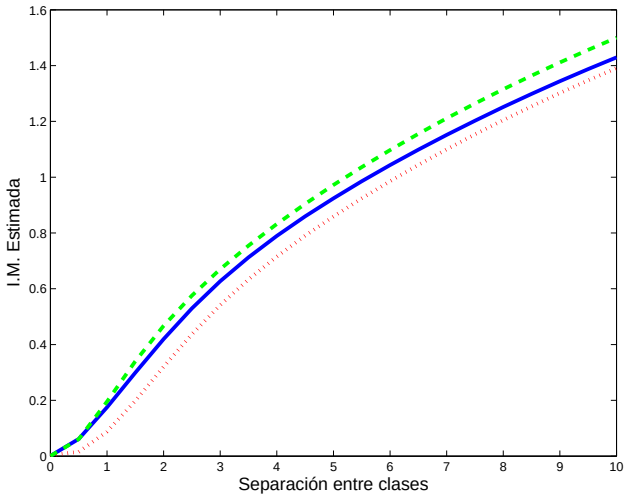

Figura 2.4: Estimación de la información mutua en un problema biclase unidimensional. El eje horizontal representa la distancia entre los centros de ambas clases. Los datos de cada clase están normalizados y tienen estadística gausiana (línea punteada), uniforme (continua) y laplaciana (discontinua).

2.3.3. Extensión a múltiples proyecciones

Hasta ahora se ha descrito el procedimiento para obtener una componente individual z_k que maximice $I(z_k, y)$. A continuación describiremos un método válido que extienda el procedimiento a la obtención de múltiples componentes como se muestra en la función de coste a maximizar en (2.3). En lugar de obtenerlas simultáneamente, se propone obtener las proyecciones secuencialemente. De esta forma conseguimos una ordenación de las proyecciones por relevancia, ya que la calidad de las mismas desciende con cada nueva proyección, al ir reduciéndose progresivamente la dimensión del subespacio de búsqueda.

En el apartado anterior se ha probado la necesidad de contar con una versión de z normalizada y centrada, para que $J(z)$ proporcione un valor preciso. Por ello establecemos la restricción $\|\mathbf{w}_k\| = 1$. De esta forma, al estar los datos blanqueados a la entrada, estarán normalizados a la salida. Ahora necesitamos imponer otra restricción: la ortogonalidad entre las proyecciones obtenidas, de forma que $\mathbf{w}_k^T \mathbf{w}_j = 0$ para $1 \leq j < i$. Hay dos razones para imponer esta restricción. En primer lugar, se evita que la misma proyección aparezca más de una vez durante el proceso. En segundo lugar, se preserva a la salida el blanqueado que se ha llevado a cabo en la entrada.

En resumen, el conjunto de proyecciones dadas por la matriz $\mathbf{W}$ han de ser ortonormales. La necesidad numérica de ortonormalidad planteada en el apartado anterior se extiende ahora por las razones expuestas. Para conseguir la ortonormalidad de forma secuencial, se debe forzar a que cada componente obtenida pertenezca al subespacio vectorial complementario al subespacio generado por las proyecciones ya obtenidas. Hay numerosas opciones para tratar con restricciones de ortogonalidad. En nuestro caso, se ha empleado un sencillo método basado en ortogonalización de Gram-Schmidt que proporciona buenos resultados. Cada iteración del algoritmo de ascenso por gradiente constará de dos pasos. El primero es el movimiento en la dirección marcada por el gradiente en (2.15) de la forma $\mathbf{w}_k^{n+1} = \mathbf{w}_k^n + \alpha \frac{\partial I(z_k^n, y)}{\partial \mathbf{w}_k^n}$. El segundo consiste obtener un $\bar{\mathbf{w}}_k^{n+1}$ de forma que sea ortogonal a las $\{\mathbf{w}_j\}_{j=1,\cdots,k-1}$, y

tenga norma unitaria

$$\boldsymbol{\varphi} = \mathbf{w}_k^{n+1} - \sum_{j<k} \langle \mathbf{w}_k^{n+1}, \mathbf{w}_j \rangle \mathbf{w}_j$$

$$\bar{\mathbf{w}}_k^{n+1} = \frac{\boldsymbol{\varphi}}{\|\boldsymbol{\varphi}\|}$$

El procedimiento descrito hasta ahora responde a una tipología incremental, en el que las proyecciones son obtenidas en orden decreciente de relevancia. Como alternativa a este esquema puede plantearse una metodología decremental. Se parte de una matriz de proyecciones sin pérdidas, por ejemplo $\mathbf{W} = \mathbf{I}$, y se busca la proyección $\mathbf{w}_1$ que minimiza $I(z_1, y)$, es decir, que extrae información irrelevante que puede ser eliminada. La matriz $\mathbf{W}$ pasa a ser la base del subespacio ortogonal a $\mathbf{w}_1$, que cumple $[\mathbf{W} \quad \mathbf{w}_1]^T[\mathbf{W} \quad \mathbf{w}_1] = \mathbf{I}$. De esta forma, la información relevante va siendo "empujada" hacia el subespacio, cada vez de menor dimensión, que permanece ortogonal a las proyecciones desechadas. Esta metodología decremental tiene la ventaja de que el número de proyecciones a obtener puede ser determinado sobre la marcha: el proceso de eliminación de proyecciones se detiene cuando se tiene constancia de que no existen más componentes no informativas. Como inconveniente, hay que destacar el hecho de que para bases de datos de alta dimensión el proceso puede ser computacionalmente muy intensivo, si la información puede ser expresada mediante un conjunto reducido de proyecciones relevantes.

2.4. Experimentos

En el presente apartado, se muestra una serie de experimentos para validar el método propuesto. Los datos usados en las simulaciones constan tanto de una base de datos generada sintéticamente como de datos reales de disponibilidad pública. Las características más relevantes de los conjuntos de datos se muestran en la Tabla 2.1.

El problema sintético fue propuesto en [Weston, 2001b] para la evaluación de

métodos de selección de características. Esta base de datos nos servirá para evaluar el comportamiento de los métodos de extracción de características frente a datos caracterizados por fronteras de clasificación con fuertes no linealidades. Para ello, las variables sintéticas son mezcladas por medio de una matriz ortonormal. La dimensión de los datos es 50. Antes de la mezcla, sólo 2 variables son relevantes. El objetivo es por tanto obtener la matriz de extracción $\mathbf{W}$ capaz de aislar esas 2 características relevantes.

El resto de bases de datos que se muestran en la Tabla 2.1 están construidas con datos reales, y han sido obtenidos del repositorio público UCI [Newman, 1998]. Se han elegido de forma que los métodos puedan ser evaluados para condiciones dispares de dimensionalidad y número de clases. En el caso de la base de datos Isolet, se ha llevado a cabo una reducción previa mediante PCA ya que la alta dimensión de los datos de entrada con respecto al número de muestras hace inviable la aplicación de algunos de los métodos de extracción de características comparados.

Datos	Entrenamiento	Test	Dimensión	Nº clases
Sintéticos	1000	500	50	2
Landsat	4435	2000	36	6
Optdigits	3823	1797	64	10
Letter	16000	4000	16	26
Isolet	6238	1559	617(40)	26

Cuadro 2.1: Características de las bases de datos evaluadas.

Los experimentos que se proponen a continuación tienen como objetivo

1. Evaluar cuantitativamente la precisión del método MMI en la estimación de la información mutua.

2. Comparar las prestaciones de las versiones incremental y decremental del algoritmo.

3. Mostrar los resultados de clasificación y compararlos con los proporcionados por otros métodos de extracción lineal de características.

4. Ilustrar la utilidad del método MMI para visualización de información relevante para clasificación.

2.4.1. Evaluación de la estimación de la información mutua

En este primer experimento, se evalúa la validez de la estimación de $\sum_k I(z_k, y)$ en lugar de $I(\mathbf{z}, y)$, en las componentes obtenidas de los datos públicos. Se mide la diferencia u *offset* entre ambas magnitudes para justificar el uso de la función de contraste propuesta en (2.3) en lugar de la propuesta en (2.2).

Además, se justifica experimentalmente el uso de PCA de acuerdo con las razones esgrimidas en el Apartado 2.2. En la Tabla 2.2, se han estimado estas magnitudes para los datos de entrada por medio del método de estimación de la información mutua propuesto en [Kraskov, 2004], y que está basado en KNN. Se ha aplicado a los conjuntos de datos públicos de la Tabla 2.1 a excepción de Isolet, a la que se ha aplicado un preprocesado artificial. Los números negativos en la tabla se deben a la imprecisión de la estimación, pero sugieren valores próximos a cero. Los resultados demuestran que el uso de PCA reduce el *offset* en los tres casos estudiados, lo que justifica su uso en el preprocesado. En la Tabla 2.3 se ha hecho lo mismo, pero con las componentes de salida cuando se aplica el método MMI. En este caso, se ha usado otro estimador de la información mutua, a partir de una aproximación a la entropía basada en modelos de Parzen optimizados de acuerdo con el método propuesto en [Leiva-Murillo, 2006], y que se detalla en el Apartado 3.2. De acuerdo con ambos estimadores, la desviación del valor de $\sum_k I(z_k, y)$ respecto al de $I(\mathbf{z}, y)$ tiene un valor lo suficientemente pequeño como para justificar el uso de $\sum_k I(z_k, y)$ como un criterio de extracción de características válido.

Datos	Sin PCA			Con PCA		
	$\hat{I}(\mathbf{x})$	$\hat{I}(\mathbf{x}\|y)$	$\hat{I}(\mathbf{x})-\hat{I}(\mathbf{x}\|y)$	$\hat{I}(\mathbf{x})$	$\hat{I}(\mathbf{x}\|y)$	$\hat{I}(\mathbf{x})-\hat{I}(\mathbf{x}\|y)$
Landsat	41.361	20.182	21.179	-0.886	-0.513	-0.373
Optdigits	15.333	8.386	6.947	3.735	3.537	0.196
Letter	23.879	19.677	4.201	10.803	7.318	3.485

Cuadro 2.2: Estimación de $I(\mathbf{x})$ e $I(\mathbf{x}|y)$ con y sin PCA previo a la entrada, mediante el método descrito en [Kraskov, 2004]. Los valores muestran que PCA reduce la redundancia a la entrada.

Datos	$\sum I(z_k, y)$	$\hat{I}(\mathbf{z})$	$\hat{I}(\mathbf{z}\|y)$	$\hat{I}(\mathbf{z}) - \hat{I}(\mathbf{z}\|y)$
	MMI	Parzen/KNN	Parzen/KNN	Parzen/KNN
Landsat	16.114	2.57/2.01	1.67/0.57	0.91 (5.6 %)/1.44(8.9 %)
Optdigits	28.526	4.28/4.68	1.85/1.24	2.43 (8.5 %)/ 3.43 (12 %)
Letter	16.925	4.79/8.29	2.59/5.91	2.19 (12 %)/2.38 (14.1 %)

Cuadro 2.3: Estimación del offset entre los valores de (1.1) y de (2.3) a la salida. Los estimadores son los basados en KNN y en modelos de Parzen. El error estimado (dos últimas columnas) contienen también el valor porcentual respecto a la primera columna.

2.4.2. MMI incremental vs. MMI decremental

En este experimento, se compara la versión incremental de MMI con la decremental. Se ha usado el conjunto de datos Landsat ya que, al no ser su dimensión muy alta, permite llevar a cabo la versión decremental con un coste computacional moderado. El valor de $I(z_k, y)$ para cada componente, según ambos métodos, se muestra en la Figura 2.5. Como se aprecia en la figura, la versión decremental obtiene proyecciones con valores más bajos de información mutua que la versión incremental. La explicación a este fenómeno reside en la imprecisión del estimador para valores bajos de información mutua en el caso decremental. Aunque las primeras proyecciones desechadas lo han sido con valores nulos, el valor real de la información mutua puede ser distinto de cero y por tanto puede estar destruyéndose información relevante. Como consecuencia de ello, las proyecciones finales arrojan valores inferiores para las $I(z_k, y)$ porque alguna información "se ha perdido por el camino".

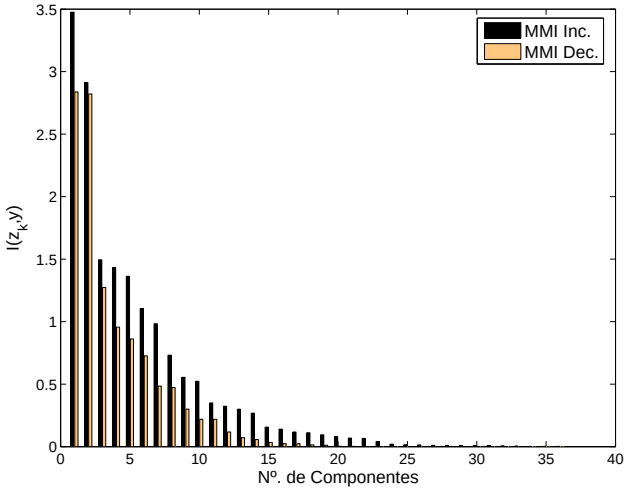

Figura 2.5: Comparación entre la información mutua obtenida por la versión incremental y la versión decremental para cada componente extraída para la base de datos Landsat.

2.4.3. Resultados de clasificación

En este último experimento, MMI (en su versión incremental) es comparado con los demás métodos sobre las bases de datos descritas en la Tabla 2.1.

Los métodos que serán evaluados a la vez que MMI son SIR, LDA, CCA (descritos en el Apartado 1.2.2) y MRMI (descrito en el Apartado 1.3.4). Dado que SIR, LDA y CCA no son capaces de obtener más de $N_c - 1$ proyecciones relevantes, se ha propuesto el siguiente esquema para poder extraer más características de estos métodos, y así poder llevar a cabo una comparación justa. Mediante un esquema basado en búsquedas en subespacios, las sucesivas proyecciones son halladas en el espacio ortogonal a las que ya han sido obtenidas. El algoritmo se muestra en la Figura 2.6.

$\mathbf{B}_1 = \mathbf{I}$
For $k = 1$ to M
    $\mathbf{x}_p = \mathbf{B}_k^T \mathbf{x}$
    Obtener proyección óptima $\tilde{\mathbf{w}}$ for $\mathbf{x}_p$ (LDA, SIR o CCA)
    Expresar en el espacio original: $\mathbf{w} = \mathbf{B}_k \tilde{\mathbf{w}}$
    $\mathbf{W}_k = \{\mathbf{W}_{k-1}, \mathbf{w}\}$
    Obtener $\mathbf{B}_{k+1}$ como la base del subespacio complementario a $\mathbf{W}_k$
end

Figura 2.6: Algoritmo para la búsqueda en subespacios complementarios. Cada proyección es hallada en el subespacio complementario a las ya obtenidas.

El preprocesado aplicado a los datos ha constado de los siguientes pasos:

- Los datos de entrenamiento son centrados mediante la rectificación de su media.
- Tras el centrado, los datos de entrenamiento son blanqueados, es decir, sus variables son decorrelacionadas mediante PCA.

- La media y la matriz de PCA obtenidas para los datos de entrenamiento son aplicadas a los datos de test.

El blanqueado de los datos permitirá una comparación más justa con MRMI. Dado que este método hace uso de un modelo de Parzen esférico para la estimación de la información mutua, una distribución esférica de los datos hace que el modelo se ajuste mejor a ellos.

Los resultados de clasificación son obtenidos mediante un clasificador KNN ($K = 1$) y un clasificador SVM con kernel RBF. Los hiperparámetros de la SVM han sido obtenidos mediante validación cruzada [Scholkopf, 2002] con 3 subconjuntos.

Los resultados, para KNN y SVM y diferentes grados de la reducción de la dimensión, se muestran en las Tablas 2.4 hasta 2.7. Se destacan, para cada columna, los valores máximos según cada uno de los dos clasificadores. La fila de "Datos en crudo" muestra, en cada tabla, el resultado de clasificación sobre la base de datos completa, sin preprocesado ni reducción en su dimensión.

El resultado más notable es el que se obtiene para la base de datos Sintética, que se muestra en la Tabla 2.4. Como puede verse, sólo MRMI y MMI son capaces de obtener la información discriminativa, aunque necesiten obtener más de dos proyecciones (la dimensión efectiva es 2). El resto de los métodos fallan dado que están basados en estadísticos de primer y segundo orden, lo que supone una incapacidad para manejar datos que exijan reglas de separabilidad no lineales. Aunque MRMI es igualmente capaz de captar información relevante, puede verse que MMI lo hace de forma más eficaz, y con sólo una proyección es capaz de extraer información discriminativa para clasificación.

Del resto de pruebas puede extraerse como conclusión importante el hecho de que, aunque MMI no es sistemáticamente mejor que los demás métodos, en los casos en que obtiene peores resultados lo hace por una mínima diferencia.

De los resultados mostrados en la Tabla 2.6 se obtiene como conclusión la alta capacidad de MMI para obtener buenos resultados con altos grados de reducción

Nº de Comps.	1	2	3	4	5
PCA/KNN	66.30	66.70	71.30	66.30	65.30
PCA/SVM	77.10	76.90	78.20	76.20	75.80
SIR/KNN	51.90	48.30	49.60	52.50	50.90
SIR/SVM	53.30	52.20	51.90	51.80	51.30
LDA/KNN	49.90	51.40	49.60	53.20	54.30
LDA/SVM	53.00	53.30	53.60	59.60	63.60
CCA/KNN	48.10	50.10	49.70	51.50	54.20
CCA/SVM	52.30	53.00	53.20	53.50	59.20
MRMI/KNN	51.50	71.50	89.00	88.10	85.50
MRMI/SVM	51.80	78.10	91.20	89.70	87.90
MMI/KNN	**93.20**	**95.50**	**94.50**	**94.00**	**91.30**
MMI/SVM	**94.90**	**96.70**	**95.70**	**94.90**	**94.10**
Datos en crudo/KNN					50.00
Datos en crudo/SVM					50.00

Cuadro 2.4: Precisión en clasificación sobre la base de datos sintética no lineal.

Nº de Comps.	1	2	3	4	5
PCA/KNN	40.80	**78.40**	**83.80**	**84.70**	**85.80**
PCA/SVM	51.55	**82.15**	85.75	**88.65**	**89.25**
SIR/KNN	47.15	71.25	82.00	83.95	82.45
SIR/SVM	55.00	79.35	85.65	86.85	87.25
LDA/KNN	47.85	71.35	82.00	83.90	82.50
LDA/SVM	54.90	79.35	85.65	86.85	87.20
CCA/KNN	47.15	71.25	82.00	83.95	82.45
CCA/SVM	55.00	79.35	85.65	86.85	87.25
MRMI/KNN	52.05	69.50	83.65	83.20	84.50
MRMI/SVM	62.15	75.75	85.95	86.40	87.10
MMI/KNN	**56.20**	74.75	82.80	84.40	84.45
MMI/SVM	**62.20**	81.40	**86.30**	87.55	87.85
Datos en Crudo/KNN					89.45
Datos en Crudo/SVM					86.85

Cuadro 2.5: Precisión sobre la base de datos Landsat.

Nº de Comps.	1	2	3	4	5	6
PCA/KNN	28.32	52.81	71.68	78.80	87.81	90.43
PCA/SVM	35.89	61.83	76.02	82.42	90.87	92.60
SIR/KNN	32.22	59.04	77.30	**85.31**	89.87	92.60
SIR/SVM	40.57	64.77	**80.80**	**88.04**	91.15	92.77
SIR/KNN	32.22	59.04	77.30	**85.31**	89.87	92.60
LDA/SVM	40.57	64.77	**80.80**	**88.04**	91.15	92.71
CCA/KNN	32.22	59.04	77.30	**85.31**	89.87	92.60
CCA/SVM	40.57	64.77	**80.80**	**88.04**	91.15	92.77
MRMI/KNN	35.45	58.82	**77.74**	23.15	24.04	25.43
MRMI/SVM	41.29	65.16	**80.80**	33.44	34.84	35.61
MMI/KNN	**36.39**	**60.32**	72.12	84.42	**89.98**	**92.65**
MMI/SVM	**42.57**	**66.17**	77.57	86.64	**92.15**	**93.43**
Datos en Crudo/KNN						98.00
Datos en Crudo/SVM						97.16

Cuadro 2.6: Precisión obtenida sobre los datos Optdigits.

Nº de Comps.	1	2	3	4	5	10
PCA/KNN	4.50	6.73	10.50	13.35	20.62	79.93
PCA/SVM	7.72	9.07	13.57	17.12	24.22	85.42
SIR/KNN	15.05	16.70	23.57	37.15	47.95	86.05
SIR/SVM	9.10	13.42	23.35	37.35	52.20	90.17
LDA/KNN	15.05	16.70	23.57	37.15	47.95	86.05
LDA/SVM	9.12	13.42	23.35	37.35	52.20	90.17
CCA/KNN	15.05	16.70	23.57	37.15	47.95	86.05
CCA/SVM	9.12	13.42	23.35	37.35	52.20	90.17
MRMI/KNN	15.15	16.20	24.68	**37.35**	47.58	**89.20**
MRMI/SVM	10.17	13.70	24.42	37.17	51.65	**92.27**
MMI/KNN	**15.18**	**16.73**	**26.20**	35.45	**48.20**	84.25
MMI/SVM	**11.50**	**14.43**	**27.83**	**37.53**	**52.82**	89.47
Datos en Crudo/KNN						95.65
Datos en Crudo/SVM						97.55

Cuadro 2.7: Precisión obtenida sobre los datos Letter.

Nº de Comps.	1	2	3	4	5	6
PCA/KNN	9.11	11.67	16.16	18.47	24.44	27.45
PCA/SVM	13.15	18.35	25.14	27.90	32.84	37.72
SIR/KNN	5.45	22.00	42.33	61.77	69.34	72.48
SIR/SVM	7.89	32.91	53.75	70.43	75.50	79.54
LDA/KNN	19.82	**40.28**	55.93	65.62	69.34	**75.95**
LDA/SVM	27.65	49.90	65.75	72.23	75.88	79.92
CCA/KNN	19.69	**40.28**	55.93	65.62	69.34	**75.95**
CCA/SVM	27.13	50.29	66.13	72.48	76.65	79.41
MRMI/KNN	21.55	38.87	**58.56**	48.49	59.33	64.08
MRMI/SVM	25.02	**50.22**	**66.84**	58.56	67.61	74.41
MMI/KNN	**21.81**	40.22	53.88	**66.00**	**72.16**	73.57
MMI/SVM	**28.74**	49.90	61.19	**74.66**	**78.77**	**80.95**
Datos en Crudo/KNN						85.31
Datos en Crudo/SVM						95.38

Cuadro 2.8: Precisión obtenida sobre los datos Isolet.

de la dimensión (para una dimensión de 1 ó 2 obtiene el mejor resultado). A la luz de su comparación con MRMI, queda clara la mejor disponibilidad de MMI para solventar el problema de la maldición de la dimensión. El hecho de que esta base de datos presente una alta dimensión de partida hace que el uso de modelos basados en ventanas de Parzen puedan padecer un sobreajuste a los datos. De ahí que MRMI se comporte bien para una baja dimensión a la salida, pero sus prestaciones se deterioran significativamente para $d > 3$, situación para la que el modelo esférico a la salida deja de describir bien los datos.

En la Tabla 2.7 queda patente la superioridad de MMI en una base de datos con un elevado número de clases. La explicación a esta superioridad puede hallarse en el hecho de que MMI puede encontrar los patrones de discriminación no lineal de forma más eficiente que los métodos clásicos, ya que la presencia de un elevado número de clases aumenta la dificultad de encontrar proyecciones que permitan la discriminación entre ellas.

2.4.4. Visualización de información discriminante

El algoritmo propuesto puede ser utilizado para encontrar las variables más informativas en un problema de clasificación. Para ello, no se lleva a cabo ninguna extracción de características, sino que se evalúa $I(x_k, y)$ directamente sobre los datos sin procesar. En un problema de reconocimiento en imagen, esto puede ser de interés dado que permite localizar las áreas de la imagen que tienen un mayor valor discriminante.

La base de datos Facedata consta de imágenes de tamaño 19×19, algunas de las cuales contienen una cara humana [Heisele, 2000]. En la Figura 2.7 se muestran algunos ejemplos de imágenes etiquetadas como positivas. En la Figura 2.8 se muestra la información mutua contenida en cada píxel acerca de la variable que se quiere inferir. Los valores cercanos al blanco indican alta información discriminante, mientras que los píxeles cercanos al negro indican valores bajos. En el mapa de la Figura 2.8 queda claro que los píxeles con mayor capacidad discriminante son los situados en las zonas

de la imagen que contienen los rasgos característicos de una cara humana: frente, ojos, pómulos y boca.

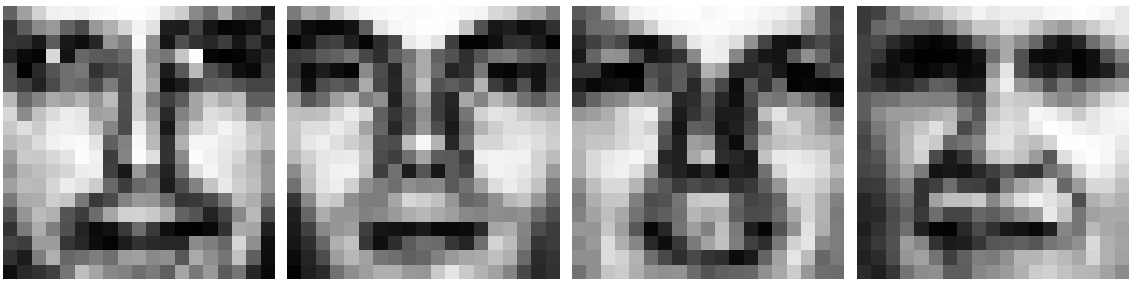

Figura 2.7: Muestras etiquetadas como positivas en la base de datos Facedata.

Figura 2.8: "Mapa de Información" proporcionado por el estimador de información mutua propuesto, para la base de datos Facedata.

Se ha llevado a cabo el mismo procedimiento a la base de datos Optdigits, que consta de imágenes 8×8 de dígitos manuscritos. Se muestran algunos ejemplos de estos dígitos en la Figura 2.9. El resultado de visualizar la información mutua entre cada píxel y el vector de clases se muestra en la Figura 2.10. Aunque en este caso el resultado es menos intuitivo, puede apreciarse cómo la información útil para discriminar entre dígitos se localiza principalmente en las zonas de la imagen que permiten

distinguir entre grupos de dígitos. El conocimiento de este "mapa de información" puede ser muy útil dado que nos puede servir para descartar desde el principio zonas de la imagen que no son relevantes para la discriminación, como los bordes izquierdo y derecho.

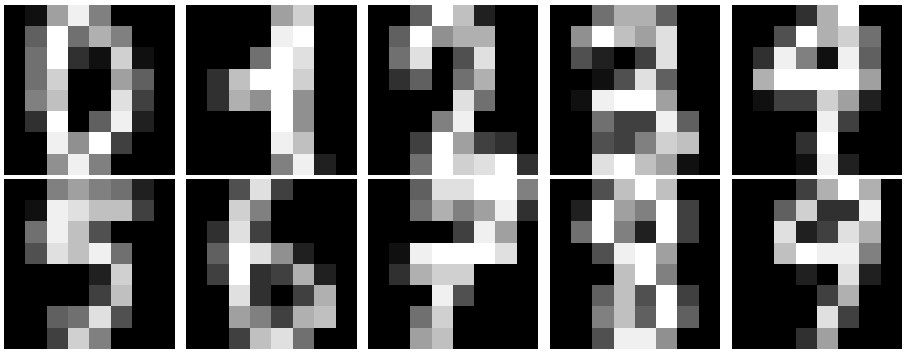

Figura 2.9: Ejemplos de imágenes tomadas de la base de datos Optdigits. Se muestra un ejemplo de cada clase: dígitos manuscritos desde el 0 hasta el 9.

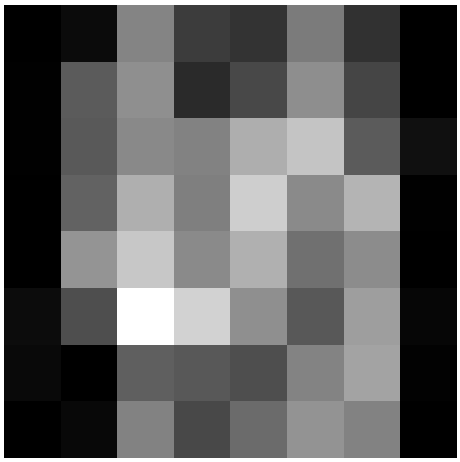

Figura 2.10: "Mapa de Información" proporcionado por el estimador de información mutua propuesto, para la base de datos Optdigits.

2.5. Valoración de los resultados y conclusiones

La principal limitación del método propuesto es su carácter incremental, lo que hace que sea subóptimo cuando se aplica a conjuntos de datos en los que una única proyección no sirve para discriminar. Un ejemplo de lo expuesto lo encontramos en las situaciones del tipo de la función XOR: cuando la etiqueta es generada a partir de la función OR exclusiva de dos de las variables, una única proyección no puede reunir la información discriminativa. Aún así, y de acuerdo con los resultados expuestos en la Tabla 2.4, el método ha sido capaz de aislar la información relevante, que en esta base de datos traza una función de discriminación de ese tipo. A pesar de las limitaciones de la implementación incremental y unidimensional de MMI, consigue superar al método MRMI, que sí es capaz de proporcionar varias proyecciones simultáneamente. En cuanto al resto de experimentos presentados, hay que destacar el hecho de que MMI mejora en general, aunque sin mucho margen, a los otros métodos. Por tanto, la principal conclusión a extraer es que MMI proporciona prestaciones similares o mejores que los métodos clásicos cuando los datos están definidos por patrones de discriminación de tipo lineal, y que dichas prestaciones pueden ser drásticamente mejoradas cuando las fronteras de clasificación son fuertemente no lineales.

Las razones por las que se ha construido el algoritmo MMI de forma secuencial, obteniéndose las proyecciones una a una, obedecen a la inexistencia de estimadores fiables de la entropía para datos multidimensionales, como se detalló en el Apartado 2.3. El trabajo expuesto en [Hulle, 2005] plantea una generalización de la expansión de Edgeworth a múltiples dimensiones. Este esquema plantea un doble problema: por un lado, la falta de robustez propia de dicha expansión; por otro, la necesidad de trabajar con tensores, lo que dispara la complejidad computacional cuando sube el número de componentes a la salida. El número de cumulantes a evaluar es de $d+2\binom{d}{2}+\binom{d}{3}$, por lo que la complejidad del estimador es de $O(Nd^3)$. En comparación, la complejidad del método MMI propuesto en el presente Capítulo es de $O(Nd)$.

En [Welling, 2005] se propone una estimación robusta de la FDP, y que supone

una interesante alternativa a las expansiones de Edgeworth y Gram-Charlier. Sin embargo, se ha de hallar la expansión que permita su aplicación a la estimación de la entropía. Encontrar dicha expansión así como su aplicación al esquema de maximización de la información mutua propuesto en este capítulo son tareas que quedan propuestas como líneas futuras de trabajo prioritarias.

Dado que el trabajo aquí presentado se basa en la estimación y maximización de la información mutua entre cada una de las características obtenidas y las clases, se necesita un mecanismo que permita medir y limitar, de forma más fiable, la diferencia entre la magnitud a maximizar $\sum_k I(z_k, y)$ y la magnitud que idealmente habría que maximizar, $I(\mathbf{z}, y)$. En este capítulo se ha propuesto usar la decorrelación PCA, además de la ortogonalización de Gram-Schmidt, para limitar la redundancia entre proyecciones extraídas. Sin embargo, dicho procedimiento no tiene en cuenta la dependencia estadística entre los datos, que puede residir en los estadísticos de orden mayor que dos. La principal dificultad para introducir la dependencia estadística entre las características obtenidas, con el objeto de penalizar la redundancia entre las mismas, es que implica trabajar con la información mutua multidimensional, que como se ha visto es un problema aún no resulto satisfactoriamente.

Capítulo 3

Optimización de modelos no paramétricos para aprendizaje

Aunque en el capítulo anterior se ha propuesto un método inspirado en teoría de la información que elude la problemática de la estimación de la FDP de los datos, muchos de los métodos propuestos en la literatura sí llevan a cabo esta estimación. La mayor parte de dichos métodos hacen uso de modelos no-paramétricos basados en ventanas de Parzen. La razón de la popularidad de estos modelos en este ámbito es su versatilidad, dado que presentan ciertas propiedades que recomiendan su uso. La primera ventaja es que no realizan ninguna suposición sobre la distribución de los datos. Estos modelos son suficientemente flexibles como para reproducir con fidelidad funciones de densidad de distinta naturaleza. En segundo lugar, permiten modelar datos multidimensionales aunque, en general, esto se produce bajo la suposición de esfericidad. Por último, estos modelos permiten una caracterización inmediata de la FDP de los datos a la salida de una transformación determinista, conocido el modelo a la entrada. Los centroides en la distribución de salida equivalen a la transformación de los de la entrada. Para el ancho óptimo de la ventana, sin embargo, no hay forma de establecer una relación entre el de entrada y el de salida, por lo que suele recurrirse a técnicas heurísticas.

En realidad, estas virtudes son compartidas con los modelos semiparamétricos basados en mezclas de gausianas ó GMM. La diferencia fundamental entre los GMM y los modelos de Parzen es que éstos últimos no necesitan un entrenamiento del modelo: éste está definido únicamente por los centroides de las ventanas, que a su vez están determinados por los datos. Sin embargo, un inconveniente de los modelos de Parzen frente a los GMM es la alta complejidad computacional de la evaluación del modelo, dado que la densidad de probabilidad en un punto ha de ser estimada tomando todas las muestras que conforman el conjunto de entrenamiento. Además, la precisión del modelo depende estrechamente del ancho de la ventana considerada. Esto implica que, a pesar de la denominación de métodos no paramétricos, la precisión de la estimación de la FDP depende estrechamente del correcto dimensionamiento del ancho de la ventana.

Ambos problemas han sido ya ampliamente tratados en la literatura. Para el problema de la reducción de la complejidad se han propuesto distintos métodos (ver, por ejemplo, [Fukunaga, 1989] y [Jeon, 1994]) y, aunque es importante considerarlo para su aplicación práctica a problemas de aprendizaje, no se tratará en la presente tesis. Para el problema de la elección óptima del ancho de ventana se han propuesto numerosos criterios de diseño. El presente capítulo está dedicado a presentar un método para el dimensionado del ancho de ventana, de forma que proporcione el modelo que permita estimar magnitudes como entropía o verosimilitud con suficiente fiabilidad.

En el siguiente apartado se describe en detalle el problema del diseño del ancho de ventana, y se lleva a cabo una revisión de los métodos que han sido propuestos para resolver este problema. Posteriormente se presentan tres versiones de un algoritmo en punto fijo para la elección óptima del ancho de ventana de acuerdo con el criterio de máxima verosimilitud. Para finalizar, se describe la estimación de la entropía a partir de los datos mediante estos modelos, y se demuestra la optimalidad de dicha estimación.

3.1. El problema del diseño de los modelos

Un modelo de densidad de probabilidad de Parzen toma la forma

$$\hat{p}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{x} - \mathbf{x}_i|\boldsymbol{\theta}) \tag{3.1}$$

donde $k(\cdot)$ es la ventana o núcleo del modelo. La diferencia entre esta expresión y la que se mostró en (1.25) (pág. 37) es que ahora se ha hecho explícita la dependencia de la ventana respecto a algún parámetro o conjunto de parámetros $\boldsymbol{\theta}$.

Algunos ejemplos de ventanas usadas en la construcción de estos modelos son la uniforme y la gausiana. El caso de la ventana uniforme o rectangular presenta el inconveniente de dar lugar a un modelo de densidad discontinuo. El caso de la ventana gausiana no presenta este problema, y tiene la virtud de proporcionar modelos continuos y derivables. La ventana gausiana es, de hecho, la más ampliamente utilizada en la literatura.

El conjunto de parámetros dado por $\boldsymbol{\theta}$ da la medida del ancho de dicha ventana. En el caso de la ventana uniforme, el único parámetro necesario para caracterizarla es su ancho. En cuanto a la ventana gausiana, el parámetro que describe su ancho es la varianza, en el caso unidimensional, o la matriz de covarianza en el caso multidimensional.

En la Figura 3.1 se muestran algunos ejemplos de modelos unidimensionales generados con ventanas gausianas y uniformes, para el mismo conjunto muestral y con distintos valores para su ancho. La elección del ancho de la ventana va a determinar el grado de precisión del modelo. De los modelos representados en la Figura 3.1 se desprenden dos hechos: una ventana demasiado amplia puede no ser capaz de capturar la estructura de la densidad, que en este ejemplo es bimodal. Por contra, una ventana demasiado estrecha sobreajustará el modelo a los datos a partir de los que se ha generado. Una situación intermedia registrará adecuadamente la bimodalidad de los datos, a la vez que presentará una "suavidad" que permita una buena capacidad de generalización del modelo probabilístico para nuevas realizaciones de la variable aleatoria.

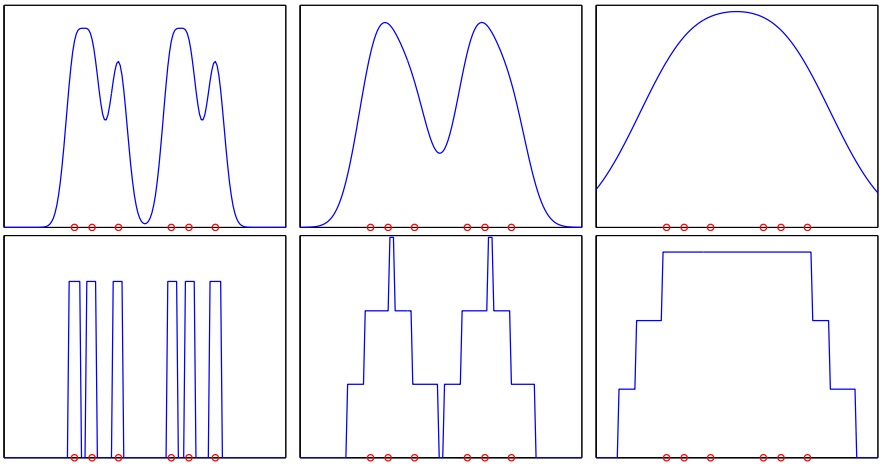

Figura 3.1: Ejemplos de modelos con ventana gausiana (arriba) y rectangular (abajo) para distintos anchos de ventana: estrecho (izquierda), moderado (centro) y amplio (derecha).

Los trabajos presentes en la literatura sobre el diseño del ancho de la ventana se han centrado principalmente en el caso unidimensional; el caso multidimensional ha sido menos tratado, y la mayor parte de las soluciones propuestas están basadas en generalizaciones del caso univariable. Los criterios para definir la idoneidad de uno u otro ancho son variados. En [Jones, 1996] se hace una revisión de los métodos propuestos en la literatura, y se sugiere una división entre métodos de diseño de primera y de segunda generación. De acuerdo con esta clasificación, algunos de los métodos de esa primera generación estarían basados en criterios de error cuadrático medio (*mean squared error*, MSE), que trata de minimizar $E\{(p(x)-\hat{p}(x))^2\}$; el error cuadrático medio integrado (*mean integrated squared error* MISE), que se basa en evaluar $E\{\int(p(x)-\hat{p}(x))^2dx\}$, y el MISE asintótico (*asymtotical MISE* ó AMISE). Este último analiza el caso $N \to \infty$, que puede estudiarse mediante una expansión de Taylor de segundo orden, y por tanto exige un conocimiento de las derivadas primera y segunda de $p(x)$ [Jones, 1996] [Silverman, 1986]. Los algoritmos de segunda generación incluirían a los métodos *plug-in*, que proporcionan una estimación del

término cuadrático en la expansión de Taylor. De esta forma se llega a una ecuación en punto fijo que minimiza el AMISE del estimador. También se incluyen en esta segunda generación los métodos basados en técnicas de remuestreo o *bootstrap*, que construyen modelos a partir de datos generados con remuestreo a partir de la distribución de datos empírica [Jones, 1996]. Otros métodos han hecho uso de la divergencia de Kullback-Leibler para medir la disparidad del estimador respecto a la densidad real [Bowman, 1984].

3.1.1. Diseño de modelos de Parzen basado en máxima verosimilitud

El criterio de máxima verosimilitud (*maximum likelihood*, ML) es el más ampliamente utilizado en el diseño de modelos paramétricos y semi-paramétricos, y se basa en obtener el modelo que mejor explica los datos de partida. En el caso de los modelos paramétricos, se parte de una familia de modelos dependientes de un conjunto de parámetros $\boldsymbol{\theta}$ que hay que ajustar. El ajuste se realiza de forma que se maximice la log-verosimilitud

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\mathbf{X}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log \hat{p}(\mathbf{x}_i|\boldsymbol{\theta}) \tag{3.2}$$

donde cada $\hat{p}(\mathbf{x}_i|\boldsymbol{\theta})$ es la verosimilitud del modelo, evaluado en $\mathbf{x}_i$, sujeto al conjunto de parámetros $\boldsymbol{\theta}$.

En el caso de los modelos semiparamétricos de mezcla de gausianas o GMM, el procedimiento es similar. Estos modelos vienen dados por una suma de L modelos paramétricos (en este caso gausianos), de la forma que se vio en (1.24) (pag. 36). Ahora, $\boldsymbol{\theta}_l$ es el conjunto de parámetros asociado al submodelo gausiano $f(\mathbf{x}|\boldsymbol{\theta}_l) = G(\mathbf{x}|\boldsymbol{\theta}_l)$, y que en este caso abarca la matriz de covarianza. El criterio ML puede ser aplicado aquí también, con la dificultad añadida de la estimación de los pesos α_l. El algoritmo EM permite la obtención simultánea de los parámetros y los pesos α_i mediante el criterio de máxima verosimilitud [Bishop, 1995].

Aunque los modelos de Parzen entran en la categoría de métodos no paramétricos, hemos podido ver que el ancho de ventana utilizado determina la precisión del modelo, por lo que puede tomar la consideración de parámetro a optimizar. Por tanto, la expresión para la verosimilitud (3.2) sigue siendo aplicable en este caso, donde $\boldsymbol{\theta}$ denota el ancho de la ventana. Como puede intuirse fácilmente, la aplicación del criterio ML a estos modelos no es posible directamente si evaluamos la verosimilitud en las muestras con las que hemos generado el modelo. Considérese un modelo de Parzen como en (3.1), con ventana gausiana y dependiente de un único parámetro $\boldsymbol{\theta} = \sigma^2$, es decir, las gausianas son esféricas, con varianza σ^2 en cada una de las dimensiones. En tal caso, el ancho de la ventana óptima en términos de verosimilitud es el nulo, que es el que arroja un valor de verosimilitud infinito, dado que estamos evaluándola sobre las muestras en las que están centradas las ventanas. Es decir

$$\lim_{\sigma\to 0}\sum_{i=1}^{N}\log\hat{p}(\mathbf{x}_i) = \lim_{\sigma\to 0}\sum_{i}\log\frac{1}{N}\sum_{j}(2\pi\sigma^2)^{-D/2}\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right) = \infty$$

La validación cruzada permite la aplicación del criterio ML ya que evaluaremos el modelo sobre muestras distintas a las que conforman el modelo. Cuando la validación cruzada se lleva a cabo de forma que se usen todas las muestras a excepción de una para construir un modelo o clasificador que se evalúa sobre la restante, el procedimiento recibe el nombre de "dejar uno fuera" (*leave-one-out*, LOO). En el caso del modelado de una FDP, esto da lugar a una estimación del tipo

$$\hat{p}(\mathbf{x}_i) = \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} G(\mathbf{x}_i-\mathbf{x}_j|\boldsymbol{\theta}) \tag{3.3}$$

donde ya hemos concretado el uso de una ventana de tipo gausiano. Este procedimiento fue propuesto por primera vez en [Duin, 1976] y ha sido posteriormente analizado por otros autores [Silverman, 1986], [Hall, 1982]. Sin embargo, los métodos presentados en la literatura para el diseño óptimo a partir de la verosimilitud y validación cruzada tienen dos importantes limitaciones. En primer lugar, se centran en el caso de variables unidimensionales. El segundo es que no cuentan con un

procedimiento cerrado de optimización, por lo que se elige el σ^2 óptimo mediante un barrido en una malla de valores de ancho de ventana, escogiéndose el valor máximo.

A continuación, presentamos un procedimiento de diseño para la obtención de $\boldsymbol{\theta}$ que supera estos dos inconvenientes. Para ello, calcularemos el gradiente $\nabla_{\boldsymbol{\theta}} \log L(\mathbf{X}|\boldsymbol{\theta})$, y lo igualaremos a cero. Veremos que esto da lugar a un algoritmo de punto fijo, que denominaremos ML-LOO, que permite optimizar $\boldsymbol{\theta}$ mediante un procedimiento iterativo.

3.2. Algoritmos de punto fijo para la optimización del modelo

A la hora de construir un modelo no paramétrico con ventanas o núcleos gausianos, cabe distinguir entre tres casos, que difieren en el grado de complejidad que asumimos para la matriz de covarianza de estas ventanas. En primer lugar, puede suponerse una forma esférica para $\mathbf{C}$, de manera que $\mathbf{C} = \sigma^2\mathbf{I}_D$, donde $\mathbf{I}_D$ es la matriz identidad $D \times D$. En el segundo caso, se permite que las gausianas tengan un ancho distinto en cada una de las dimensiones, y la matriz de covarianza tiene una forma diagonal, de forma que $\sigma_{ij} = 0$ si $i \neq j$. En el tercer caso, no habrá restricciones sobre la forma de $\mathbf{C}$, pudiendo ésta recoger libremente el escalado de los datos en las distintas dimensiones así como la interacción entre las componentes.

3.2.1. El caso esférico

El uso de una matriz de covarianza esférica es equivalente a modelar las distintas componentes de forma separada, con el mismo ancho. En este caso, $\boldsymbol{\theta} = \sigma^2$. El valor de una gausiana entonces viene dado por

$$\begin{aligned} G_{ij}(\sigma^2) &= G(\mathbf{x}_i - \mathbf{x}_j|\sigma^2) \\ &= (2\pi)^{-D/2}\sigma^{-D} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2 \right) \end{aligned}$$

Y el valor de su gradiente respecto a σ es

$$\nabla_\sigma G_{ij}(\sigma^2) = \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^3} - \frac{D}{\sigma}\right) G_{ij}(\sigma^2)$$

Entonces, la derivada de la log-verosimilitud, calculada como para los modelos paramétricos en (3.2), es

$$\begin{aligned}\nabla_\sigma \log L(\mathbf{X}|\sigma^2) &= \sum_i \frac{1}{\hat{p}(\mathbf{x}_i)} \frac{1}{N-1} \sum_{j\neq i} \frac{\partial}{\partial\sigma} G_{ij}(\sigma^2) \\ &= \frac{1}{N-1} \sum_i \frac{1}{\hat{p}(\mathbf{x}_i)} \sum_{j\neq i} \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^3} - \frac{D}{\sigma}\right) G_{ij}(\sigma^2)\end{aligned}$$

Buscamos ahora el máximo de $\log L(\mathbf{X}|\sigma^2)$, y por tanto el punto en el que se anula su derivada. Obtenemos entonces

$$\begin{aligned}\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)} \sum_{j\neq i} \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^3} G_{ij}(\sigma^2) &= \sum_i \frac{1}{\hat{p}(\mathbf{x}_i)} \frac{D}{\sigma} \sum_{j\neq i} G_{ij}(\sigma^2) \\ &= \frac{N(N-1)D}{\sigma}\end{aligned}$$

Para la segunda igualdad se ha usado el hecho de que, por definición, $\sum_{j\neq i} G_{ij} = (N-1)\hat{p}(\mathbf{x}_i)$. Tras aislar la σ^2, obtenemos el siguiente algoritmo de punto fijo

$$\sigma^2_{n+1} = \frac{1}{N(N-1)D} \sum_i \frac{1}{\hat{p}_n(\mathbf{x}_i)} \sum_{j\neq i} \|\mathbf{x}_i - \mathbf{x}_j\|^2 G_{ij}(\sigma^2_n) \tag{3.4}$$

donde $\hat{p}_n$ denota el modelo de Parzen obtenido en la iteración n (es decir, aquel que hace uso del ancho dado por σ^2_n).

A continuación se estudia la convergencia del algoritmo, definida a partir del siguiente teorema:

Teorema 3.1 *El algoritmo en* (3.4) *tiene un punto fijo en el intervalo dado por* $\left(\frac{\overline{d^2_{NN}}}{D}, \frac{2\,\mathrm{tr}\{\mathbf{C_x}\}}{D}\right)$, *siendo* $\overline{d^2_{NN}}$ *la distancia cuadrática media de las muestras a su vecino más próximo, y* $\mathbf{C_x}$ *la matriz de covarianza de* $\mathbf{x}$. *Además, el punto fijo es único y el algoritmo tiene garantizada su convergencia en dicho intervalo si se cumple*

$$\frac{1}{2\sigma^4 N(N-1)^2 D} \sum_i \frac{1}{\hat{p}_l(\mathbf{x}_i)^2} \sum_{j\neq i} \sum_{k\neq i,j} (d^2_{ij} - d^2_{ik})^2 \exp\left(-\frac{d^2_{ij} + d^2_{ik}}{2\sigma^2}\right) < 1 \tag{3.5}$$

Demostración Sea $\sigma^2 = g(\sigma^2)$ el algoritmo en (3.4). La demostración de la existencia del punto fijo se basa en constatar la existencia de un intervalo en el que se cumple $a < g(\sigma^2) < b$ si $\sigma^2 \in (a, b)$.

Para demostrar que el intervalo $\left(\frac{\overline{d^2_{NN}}}{D}, \frac{2\,\mathrm{tr}\{\mathbf{C_x}\}}{D}\right)$ cumple esa condición, necesitamos demostrar estos tres hechos:

1. $g\left(\frac{\overline{d^2_{NN}}}{D}\right) > \frac{\overline{d^2_{NN}}}{D}$
2. $g\left(\frac{2\,\mathrm{tr}\{\mathbf{C_x}\}}{D}\right) < \frac{2\,\mathrm{tr}\{\mathbf{C_x}\}}{D}$
3. $g(\sigma^2)$ es monótona creciente en el intervalo con σ^2.

De esta forma se garantiza al menos un punto de corte entre la función $g(\sigma^2)$ y la recta $g(\sigma^2) = \sigma^2$ Para la demostración del primer punto reescribimos (3.4) de la forma

$$g(\sigma^2) = \frac{1}{ND} \sum_i \sum_{j \neq i} d^2_{ij} \frac{1}{1 + \sum_{k \neq i,j} \exp(\frac{d^2_{ij} - d^2_{ik}}{2\sigma^2})} \tag{3.6}$$

El límite en cero viene dado por

$$\lim_{\sigma^2 \to 0} f(\sigma^2) = \frac{1}{ND} \sum_i \min_{j \neq i} d^2_{ij} = \frac{\overline{d^2_{NN}}}{D}$$

porque el denominador en (3.6) se hace nulo (alguna exponencial tiende al infinito) salvo en los casos en que $d^2_{ij} < d^2_{ik}$, $\forall k \neq j$.

Asumida $g(\sigma^2)$ monótona creciente, el primer punto queda demostrado ya que $\frac{\overline{d^2_{NN}}}{D}$ es el valor mínimo que puede tomar $g(\sigma^2)$.

Para demostrar el segundo punto, tomamos el límite en el infinito

$$\begin{aligned} \lim_{\sigma^2 \to \infty} g(\sigma^2) &= \lim_{\sigma^2 \to \infty} \frac{1}{ND} \sum_i \sum_{j \neq i} d^2_{ij} \frac{\exp(-\frac{d^2_{ij}}{2\sigma^2})}{\sum_{k \neq i} \exp(-\frac{d^2_{ik}}{2\sigma^2})} \\ &= \frac{1}{ND} \sum_i \sum_{j \neq i} d^2_{ij} \frac{1}{N-1} \end{aligned}$$

El sumatorio de distancias puede ser expresado en términos del operador esperanza, de la forma

$$\frac{1}{N(N-1)}\sum_i\sum_{j\neq i} d_{ij}^2 = E_{i,j}\{(\mathbf{x}_i-\mathbf{x}_j)^T(\mathbf{x}_i-\mathbf{x}_j)\} = 2E_i\{\mathbf{x}_i^T\mathbf{x}_i\} - 2\boldsymbol{\mu_x}^T\boldsymbol{\mu_x}$$

Una propiedad del álgebra lineal es que si $\boldsymbol{\mu_x} = E_{\mathbf{x}}\{\mathbf{x}\}$ y $\mathbf{C_x} = E_{\mathbf{x}}\{\mathbf{x}\mathbf{x}^T\} - \boldsymbol{\mu_x}\boldsymbol{\mu_x}^T$, entonces $E\{\mathbf{x}^T\mathbf{x}\} = \mathrm{tr}\{\mathbf{C_x}\} + \boldsymbol{\mu_x}^T\boldsymbol{\mu_x}$. De acuerdo con esto, tenemos

$$2E\{\mathbf{x}^T\mathbf{x}\} - 2\boldsymbol{\mu_x}^T\boldsymbol{\mu_x} = 2\,\mathrm{tr}\{\mathbf{C_x}\}$$

De esta forma, queda

$$\lim_{\sigma^2\to\infty} g(\sigma^2) = \frac{2\,\mathrm{tr}\{\mathbf{C_x}\}}{D}$$

El segundo punto queda demostrado ya que el valor máximo de $g(\sigma^2)$ tiene lugar en el infinito.

Para verificar el tercer punto calculamos la derivada y comprobamos que es positiva

$$\begin{aligned}
\frac{dg(\sigma^2)}{d\sigma^2} &= \frac{1}{2\sigma^4 ND}\sum_i\sum_{j\neq i} d_{ij}^2 \frac{\sum_k (d_{ij}^2 - d_{ik}^2)\exp(-\frac{d_{ij}^2+d_{ik}^2}{2\sigma^2})}{(\sum_{l\neq i}\exp(-\frac{d_{il}^2}{2\sigma^2}))^2} \\
&= \frac{1}{2\sigma^4 ND}\sum_i \frac{1}{(\sum_{l\neq i}\exp(-\frac{d_{il}^2}{2\sigma^2}))^2}\sum_{j\neq i}\sum_{k\neq i} d_{ij}^2(d_{ij}^2-d_{ik}^2)\exp(-\frac{d_{ij}^2+d_{ik}^2}{2\sigma^2}) \\
&= \frac{1}{2\sigma^4 N(N-1)^2 D}\sum_i \frac{1}{\hat{p}_l(\mathbf{x}_i)^2}\sum_{j\neq i}\sum_{k\neq i,j}(d_{ij}^2-d_{ik}^2)^2\exp(-\frac{d_{ij}^2+d_{ik}^2}{2\sigma^2}) \geq 0
\end{aligned} \tag{3.7}$$

Para el último paso, se han eliminado los términos con $j = k$, que son nulos, y se han reagrupado los restantes.

Para la demostración de la convergencia del algoritmo, necesitamos comprobar que se cumpla que $|g'(\sigma^2)| < 1$ [Fletcher, 1995]. Cuando su valor es menor que uno, tenemos garantizado el hecho de que sólo existe un punto de corte entre la función $g(\sigma^2)$ y la recta $g(\sigma^2) = \sigma^2$. La condición de convergencia en (3.5) equivale a exigir que el valor de (3.7) sea menor que 1. ∎

3.2.2. El caso diagonal

En el caso de una matriz de covarianza diagonal, los elementos que no pertenecen a la diagonal son nulas

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{pmatrix} \tag{3.8}$$

Ahora el modelo es flexible al rango de los datos en sus distintas dimensiones, por lo que puede aplicarse a datos heterogéneos cuyas variables tienen una naturaleza estadística muy dispar. Pero, al igual que en el caso esférico, se sigue asumiendo independencia estadística entre sus distintas dimensiones.

El conjunto de parámetros es entonces de la forma $\boldsymbol{\theta} = \boldsymbol{\sigma}^2$, un vector D-dimensional cada uno de cuyos elementos es σ_k^2, el ancho de la ventana en la dimensión k. Se cumple que

$$G_{ij}(\boldsymbol{\sigma}^2) = G(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\sigma}^2) = \prod_{k=1}^{D} G(x_i^{(k)} - x_j^{(k)}|\sigma_k^2) = \prod_k G_{ij}^{(k)}(\sigma_k^2)$$

El gradiente de cada ventana respecto a cada σ_k^2 viene dado por

$$\nabla_{\sigma_k^2} G_{ij}(\boldsymbol{\sigma}^2) = \nabla_{\sigma_k^2} G_{ij}^{(k)}(\sigma_k^2) \times \prod_{l \neq k} G_{ij}^{(l)}(\sigma_k^2) = \big(\frac{x_i^{(k)} - x_j^{(k)}}{\sigma_k^3} - \frac{1}{\sigma_k}\big) G_{ij}(\boldsymbol{\sigma}^2)$$

Siguiendo los mismos pasos que en el apartado anterior para maximizar la verosimilitud, se llega al siguiente algoritmo de punto fijo

$$\boldsymbol{\sigma}_{n+1}^2 = \frac{1}{N(N-1)} \sum_i \frac{1}{\hat{p}_n(x_i)} \sum_{j \neq i} (\mathbf{x}_i - \mathbf{x}_j)^2 G_{ij}(\boldsymbol{\sigma}_n^2) \tag{3.9}$$

donde, abusando de notación, la potencia cuadrática del vector denota un vector con cada elemento elevado al cuadrado.

La demostración de las condiciones de existencia de un punto fijo así como de la convergencia al mismo es, con mucho, más complejo que en el caso del modelo esférico.

Esto se debe a la imposibilidad de desacoplar el algoritmo para cada elemento de la matriz. De otro modo, el procedimiento seguido en la demostración del Teorema 3.1 no es viable en este caso ya que el análisis con límites llevado a cabo anteriormente no procede con vectores.

3.2.3. El caso completo

La expresión general para una ventana gausiana con matriz de covarianza $\mathbf{C}$ es

$$G_{ij}(\mathbf{C}) = |2\pi\mathbf{C}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right)$$

y su derivada respecto a $\mathbf{C}$

$$\nabla_{\mathbf{C}} G_{ij}(\mathbf{C}) = \frac{1}{2}\left(\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T - \mathbf{I}\right)\mathbf{C}^{-1}G_{ij}(\mathbf{C})$$

Al igual que en el caso esférico, tomamos la derivada de la log-verosimilitud y la igualamos a cero

$$\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{2}\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{C}^{-1}G_{ij} = \sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{2}\mathbf{C}^{-1}G_{ij}$$

Multiplicando ambos miembros por $\mathbf{C}$, tanto por la izquierda como por la derecha, tenemos

$$\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T G_{ij} = \mathbf{C}\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i} G_{ij}$$

Aplicando las simplificaciones similares a las vistas para el caso esférico, queda

$$\sum_i \frac{1}{\hat{p}(\mathbf{x}_i)}\sum_{j\neq i}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T G_{ij} = \mathbf{C}N(N-1)$$

lo que da lugar al algoritmo de punto fijo

$$\mathbf{C}_{n+1} = \frac{1}{N(N-1)}\sum_i \frac{1}{\hat{p}_n(\mathbf{x}_i)}\sum_{j\neq i}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T G_{ij}(\mathbf{C}_n) \tag{3.10}$$

Al igual que el correspondiente al caso diagonal, este algoritmo cabe ser interpretado como una generalización del proporcionado en (3.4), ya que no se proporciona demostración de su convergencia.

Una matriz de covarianza completa proporciona modelos más versátiles que el caso esférico, por lo que se alcanzan valores de verosimilitud más altos. Como contrapartida, sufre el sobreajuste a los datos de forma más intensa que en el caso del modelo esférico, como se constatará en los resultados experimentales. Esto se debe a que en el caso completo el número de parámetros a ajustar son los elementos de la matriz $\mathbf{C}$, mientras que en el modelo esférico sólo hay que ajustar el parámetro σ^2. La matriz diagonal supone una alternativa intermedia entre el modelo esférico y el completo, y también sufre el riesgo de sobreajustar a los datos si el número de variables es reducido cuando se compara con el número de muestras.

3.3. Estimación de la entropía

El problema de la estimación de $h(\mathbf{x})$ a partir del modelo optimizado de acuerdo con las distintas versiones de ML-LOO es importante de cara a su aplicación a algunas técnicas de aprendizaje que se presentarán en el Capítulo 4. La expresión para la entropía diferencial se mostró en (1.15). Desgraciadamente, aunque dispongamos de la expresión analítica de $\hat{p}(\mathbf{x})$, la estimación de la entropía es imposible. La suma en el interior del logaritmo hace que la resolución de la integral sea impracticable analíticamente y aún difícil numéricamente [Beirlant, 1997]. Para sortear esta dificultad, se hace uso de la siguiente estimación de la entropía, que fue propuesta por primera vez en [Ahmad, 1976]

$$\hat{h}(\mathbf{x}) \approx -\frac{1}{N}\sum_{i=1}^{N}\log p(\mathbf{x}_i) \tag{3.11}$$

Nótese que, si esta estimación es llevada a cabo a partir de la densidad real $p(\mathbf{x})$, se cumple que $\hat{h}(\mathbf{x}) \to h(\mathbf{x})$ si $N \to \infty$, por la propiedad de equipartición asintótica [Cover, 2006]. Si en lugar de ello se computa sobre la densidad modelada $\hat{p}(\mathbf{x})$, se ha

demostrado la consistencia asintótica del estimador [Ahmad, 1976] y se han estudiado su sesgo y varianza [Joe, 1989]. Para distinguirla de ésta, denotaremos con $\hat{h}_{-1}(\mathbf{x})$ a la estimación LOO de la entropía, es decir, la que tiene lugar sustituyendo la densidad $p(\mathbf{x})$ en (3.11) por su estimador LOO de (3.3).

A continuación justificamos el empleo de los estimadores dados por los algoritmos en (3.4), (3.9) y (3.10) a través del enunciado del siguiente teorema:

Teorema 3.2 *De entre todas las posibles elecciones de* σ^2 *(modelo esférico),* $\boldsymbol{\sigma}^2$ *(modelo diagonal) o* $\mathbf{C}$ *(modelo completo), las proporcionadas por* (3.4)*,* (3.9) *y* (3.10) *son las que dan lugar, en media, a la estimación del tipo* $\hat{h}_{-1}$ *más cercana a la entropía real.*

Demostración El error medio cometido por el uso del modelo de Parzen, respecto al que se hubiera medido mediante la densidad real $p(\mathbf{x})$ es

$$\begin{aligned} E_{p(\mathbf{x})}\{\hat{h}_{-1}(\mathbf{x}) - h(\mathbf{x})\} &= E_{p(\mathbf{x})}\Big\{\frac{-1}{N}\sum_{i=1}^{N}\log\hat{p}(\mathbf{x}_i) - \frac{-1}{N}\sum_{i=1}^{N}\log p(\mathbf{x}_i)\Big\} \\ &= \frac{1}{N}\sum_{i=1}^{N} E_{p(\mathbf{x})}\{\log p(\mathbf{x}_i) - \log\hat{p}(\mathbf{x}_i)\} = D_{KL}(p,\hat{p}) \end{aligned}$$

La divergencia D_{KL} es una medida no negativa, y por ello llegamos al resultado

$$E_{p(\mathbf{x})}\{\hat{h}_{-1}(\mathbf{x}) - h(\mathbf{x})\} \geq 0 \tag{3.12}$$

lo que equivale a afirmar que cualquier estimador de la entropía arrojará un valor, en media, superior al real.

Por otro lado, la equivalencia entre verosimilitud y entropía, de acuerdo con (3.2) y (3.11), nos dice que

$$\hat{h}(\mathbf{x}) = -\frac{1}{N}\,\mathrm{logL}(\mathbf{x}) \tag{3.13}$$

Los algoritmos de las expresiones en (3.4),(3.9) y (3.10) han dado lugar a valores de verosimilitud máximos dentro de las familias de modelos esféricos, diagonales y completos, respectivamente, cuando la estimación de la verosimilitud se lleva a

cabo mediante LOO. De (3.13) se deduce que estos valores máximos se traducen en valores mínimos de entropía. El teorema queda inmediatamente demostrado dado que el valor de menor entropía es, conforme a la desigualdad en (3.12), el más cercano a la entropía real. ∎

En la Figura 3.2 se representan algunos ejemplos de estimaciones de la entropía en función del ancho de ventana, para datos generados por gausianas en 2 y 10 dimensiones. Queda claro a partir de la imagen el hecho de que el método propuesto proporciona el punto más cercano a la entropía real de esa familia de estimadores.

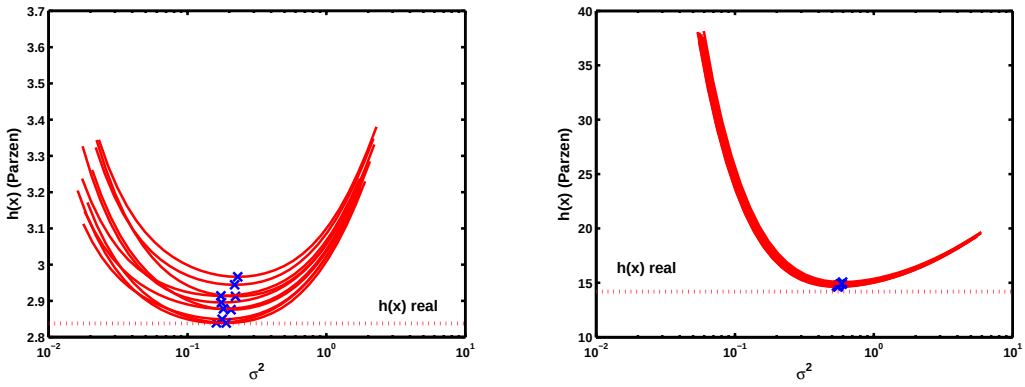

Figura 3.2: Ancho de ventana frente a entropía estimada para modelos construidos sobre 10 realizaciones de 400 muestras de una distribución normal en 2 (izquierda) y 10 (derecha) dimensiones. Se marcan los puntos escogidos por el estimador propuesto.

En la Figura 3.3 se muestran 3 ejemplos de variables bidimensionales muestreadas. La primera es una gausiana con matriz de covarianza $\mathbf{C_x} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. La segunda es una uniforme girada 45º. La tercera es una laplaciana bidimensional. Para cada una se muestra la estimación de la entropía en función del número de muestras, cuando es escogido el ancho de acuerdo con el algoritmo esférico propuesto. Se ha promediado el resultado de 20 experimentos. Puede verse cómo la precisión en la estimación crece con el número de muestras usadas, y la varianza se reduce.

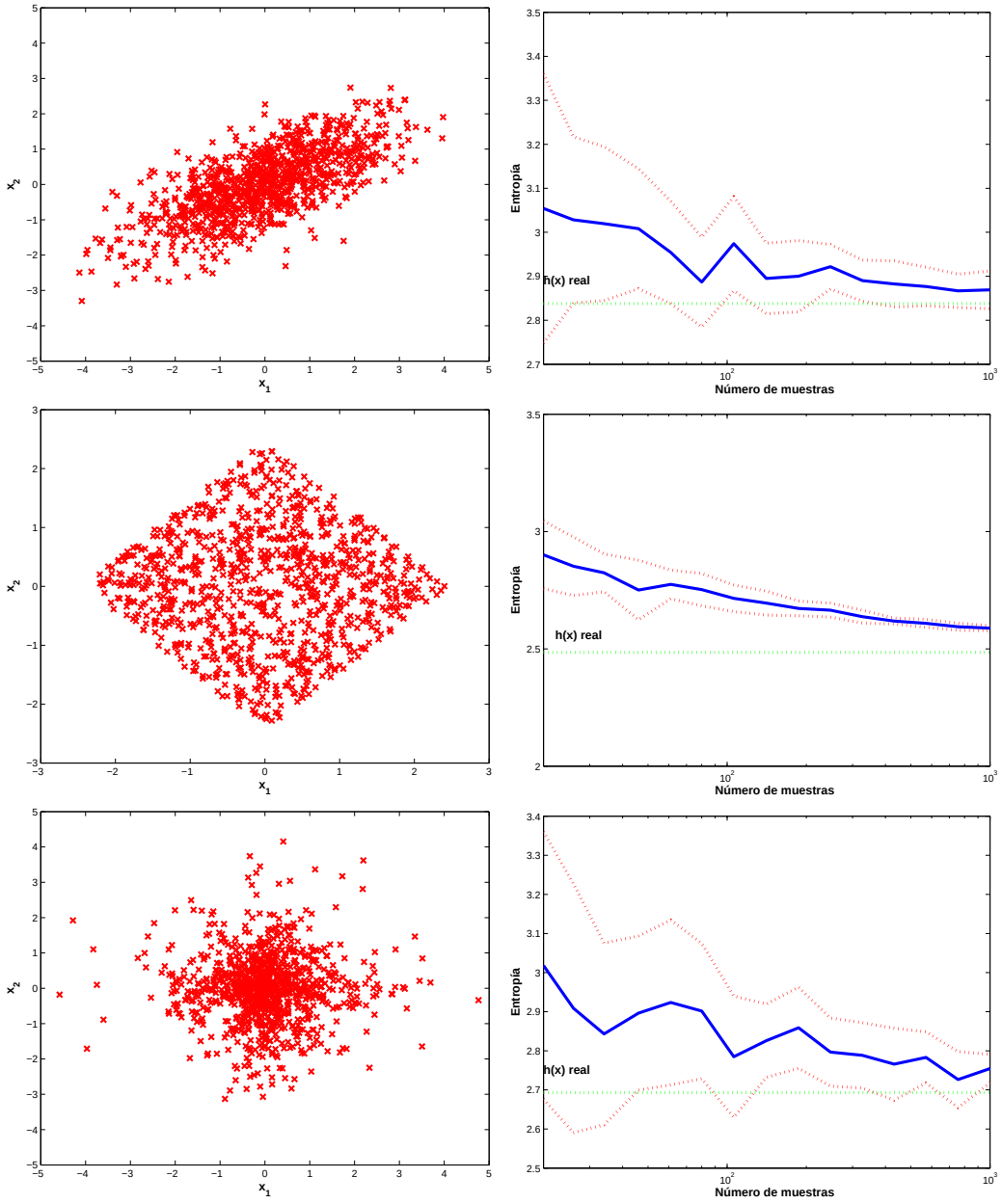

Figura 3.3: Izquierda: ejemplos de realizaciones de distribuciones gausiana, uniforme y laplaciana en dos dimensiones, respectivamente de arriba abajo. Derecha: estimaciones de la entropía muestreada en función del número de muestras utilizadas, promediadas para 100 realizaciones. Línea sólida: valor medio de las medidas; línea discontinua: margen dado por la desviación típica.

Capítulo 4

Aprendizaje basado en modelos no paramétricos optimizados

En el Capítulo 3 se ha presentado un método para optimizar modelos de Parzen de cara a su aplicación a métodos de aprendizaje que requieran técnicas de teoría de la información. El Teorema 3.2 establece que la mejor estimación de la entropía que puede hacerse con estos modelos mediante validación cruzada viene dada por el ancho de ventana obtenido por el método ML-LOO.

En el presente Capítulo se presentan una serie de aplicaciones de los modelos de Parzen, optimizados con ML-LOO, a distintos problemas de aprendizaje, con especial énfasis en la extracción lineal de características, que constituye el eje central de la tesis. En el Apartado 4.1 se muestran los resultados de aplicar los modelos optimizados, en sus distintos grados de complejidad (matriz de covarianza esférica, diagonal o completa), al problema de la clasificación de Parzen. Aunque este método de clasificación no es ampliamente utilizado en la literatura, por ser difícil el diseño adecuado del ancho de ventana, especialmente con datos multidimensionales, se mostrará cómo la solución aportada por el método ML-LOO aumenta considerablemente las prestaciones de estos clasificadores.

En el Apartado 4.2 se describe un procedimiento de extracción de características

que hace uso de los modelos optimizados. El método propuesto hace uso de teoría de la decisión para plantear como función a optimizar el test de hipótesis estimado sobre los datos de entrenamiento, modeladas estas hipótesis con ventanas de Parzen. Junto a este método, se propone otro que hace uso de modelos GMM en lugar de ventanas de Parzen, y de un criterio de máxima información mutua.

Por último, en el Apartado 4.3 se presenta un método de análisis de componentes independientes que hace uso de los modelos de Parzen optimizados.

4.1. Aplicación a clasificadores de Parzen

Un clasificador de Parzen asigna a una muestra $\mathbf{x}$ la clase dada por un criterio ML sobre los modelos construidos para cada clase, de la forma

$$\hat{y} = \arg \max_{l} \hat{p}(\mathbf{x}|c_l) \tag{4.1}$$

donde $\hat{p}(\mathbf{x}|c_l)$ es el modelo de Parzen construido con las muestras de entrenamiento de la clase l [Fukunaga, 1990].

Hay dos importantes razones por las que este clasificador no goza de mucha popularidad. La primera de ellas es que si el número de ventanas con que está construido cada modelo $\hat{p}(\mathbf{x}|c_l)$ es elevado, la complejidad computacional de la clasificación puede ser muy alta, cuando se compara con otros clasificadores. La segunda es que la precisión del clasificador depende muy directamente del dimensionado del ancho de las ventanas, ya que determina la validez de los modelos con que se lleva a cabo la clasificación.

En este apartado se pretende evaluar la validez del procedimiento presentado en el Capítulo 3 para la elección del ancho de las ventanas, mediante su incidencia en la precisión alcanzada en los clasificadores de Parzen, cuando estos están optimizados mediante el algoritmo ML-LOO. A continuación se muestra la conexión entre el criterio de clasificación de Parzen y el criterio de máxima información mutua cuando ésta es estimada mediante modelos de Parzen.

4.1.1. Información mutua, verosimilitud y clasificación de Parzen

Sea el clasificador basado en información mutua que opera de la siguiente forma: se asigna a una nueva muestra la clase de acuerdo con la cual la información mutua global es máxima. Es decir, dado un conjunto de entrenamiento $\mathbf{X}$ y un vector de etiquetas $\mathbf{Y}$, resulta razonable asignar a una nueva muestra $\mathbf{x}'$ la clase de acuerdo con la cual se maximiza el criterio dado por

$$\hat{y} = \arg\max_l \hat{I}\left(\mathbf{X}', \mathbf{Y}'_l\right) \tag{4.2}$$

donde $\hat{I}$ es un estimador muestral de la información mutua, $\mathbf{X}'$ es la unión del conjunto de entrenamiento y la muestra de test, es decir $\mathbf{X}' = \{\mathbf{X} \quad \mathbf{x}'\}$, e $\mathbf{Y}'_l$ la unión del vector de etiquetas con la clase c_l que se supone para $\mathbf{x}'$, es decir $\mathbf{Y}'_l = \{\mathbf{Y} \quad c_l\}$.

La expresión en (4.2) puede ser desarrollada de la forma

$$\begin{aligned} \hat{y} &= \arg\max_l \left[\hat{h}(\mathbf{X}') - \sum_j P_y(c_j)\hat{h}(\mathbf{X}'|\mathbf{Y}'_l = c_j)\right] \\ &= \arg\max_l \left[\hat{h}(\mathbf{X}') - \sum_{j\neq l} P_y(c_j)\hat{h}(\mathbf{X}|\mathbf{Y} = c_j) - P_y(c_l)\hat{h}(\mathbf{X}'|\mathbf{Y}'_l = c_l)\right] \end{aligned} \tag{4.3}$$

donde el término $\hat{h}(\mathbf{X}'|\mathbf{Y}'_l = c_l)$ es la entropía empírica calculada a partir de los datos de entrenamiento de la clase c_l a los que se le añade la nueva muestra. Dicho término puede ser descompuesto como

$$\begin{aligned} \hat{h}(\mathbf{X}'|\mathbf{Y}'_l = l) &= \frac{-1}{N_l + 1} \sum_{i/y_i = c_l} \log \hat{p}_l(\mathbf{x}_i) + \frac{-1}{N_l + 1} \log \hat{p}_l(\mathbf{x}') \\ &\approx \hat{h}(\mathbf{X}|\mathbf{Y}) - \frac{1}{N_l + 1} \log \hat{p}_l(\mathbf{x}') \end{aligned} \tag{4.4}$$

La última aproximación ha tenido en cuenta el hecho de que, en una situación asintótica, con un número de muestras tendiendo a infinito, se cumple que $N_l + 1 \approx N_l$. Entonces queda clara la equivalencia entre (4.3) y

$$\begin{aligned} \hat{y} &= \arg\max_l \left[\hat{h}(\mathbf{X}') - \sum_j P_y(c_j)\hat{h}(\mathbf{X}|\mathbf{Y} = c_j) + \frac{1}{N+1} \log \hat{p}_l(\mathbf{x}')\right] \\ &= \arg\max_l \left[I(\mathbf{X}, \mathbf{Y}) + C + \frac{1}{N+1} \log \hat{p}_l(\mathbf{x}')\right] \end{aligned} \tag{4.5}$$

donde el último paso está justificado por el hecho de que las magnitudes $\hat{h}(\mathbf{X}')$ y $\hat{h}(\mathbf{X})$ son constantes y su diferencia es C. Dado que $I(\mathbf{X}, \mathbf{Y})$ es también constante, podemos eliminarla del valor a maximizar. La monotonicidad del logaritmo nos lleva a

$$\begin{aligned} \hat{y} &= \arg \max_l \frac{1}{N+1} \log \hat{p}_l(\mathbf{x}') \\ &= \arg \max_l \hat{p}_l(\mathbf{x}') \end{aligned} \tag{4.6}$$

El criterio en (4.6) corresponde a un test de verosimilitudes. Cuando la estimación de la densidad $\hat{p}_l(\mathbf{x})$ corresponde a un modelo de Parzen, el criterio de clasificación da lugar al conocido clasificador de Parzen. La equivalencia entre esta clasificación y la basada en máxima información mutua queda así puesta de manifiesto. Además, la precisión del clasificador de Parzen depende estrechamente de la apropiada estimación de las entropías $\hat{h}(\cdot|y = c_l)$.

Se ha de hacer hincapié en el hecho del carácter asintótico de la demostración, ya que de otro modo la aproximación en (4.4) dejaría de tener sentido.

4.1.2. Clasificadores de Parzen mediante modelos esféricos, diagonales, completos e híbridos

A la hora de elegir entre una matriz de covarianza esférica, diagonal o completa para construir los modelos y llevar a cabo la clasificación, hay que alcanzar un compromiso entre las ventajas y debilidades de cada opción. En el caso esférico, el modelo es impreciso si las muestras no están distribuidas de una forma cercana a la esférica. Si las distintas componentes presentan una alta disparidad en cuanto a su dominio y su varianza, tomar un ancho idéntico para cada dimensión representa una suposición demasiado fuerte. Por otro lado, en el caso de un modelo completo, hay que tener en cuenta que el modelo se sobreajustará fuertemente a los datos si el número de parámetros libres es del mismo orden de magnitud que el número de muestras. El número de parámetros en el caso de un modelo de Parzen completo es $\frac{1}{2}D(D+1)$, que es el número de elementos de la matriz $\mathbf{C}$, asumido el hecho de que

$c_{ij} = c_{ji}$. El modelo diagonal se queda, por tanto, a medio camino entre uno y otro a este respecto, ya que tiene un número de parámetros igual a D.

En la Tabla 4.1 se muestran los resultados de clasificación de Parzen de acuerdo con los modelos esférico, diagonal, completo (optimizados con los algoritmos en (3.4), (3.9) y (3.10)), y el que viene dado por la regla de Scott, que establece una matriz de covarianza para la ventana de la forma $\mathbf{C} = N^{\frac{-2}{D+4}}\hat{\mathbf{C}}_{\mathbf{x}}$, donde $\hat{\mathbf{C}}_{\mathbf{x}}$ es la matriz de covarianza empírica de $\mathbf{x}$ [Scott, 1992]. Junto a estos resultados, se incluyen los obtenidos por los métodos de clasificación discriminativa KNN y SVM. Los resultados para Pima y Wine han sido obtenidos mediante *leave-one-out*, dado que no se proporcionan datos de entrenamiento y test por separado.

Como puede verse, el clasificador de Parzen optimizado ofrece, para las tres versiones, resultados competitivos incluso cuando es comparado con métodos discriminativos puros como KNN o SVM. Al ser el clasificador propuesto un método generativo, tiene la ventaja sobre aquellos de proporcionar una salida que es interpretable en términos probabilísticos.

Los resultados muestran el efecto del sobreajuste al que se aludía anteriormente. El modelo esférico proporciona, en general, mejores prestaciones que el modelo diagonal y que el completo, especialmente en las bases de datos de alta dimensionalidad o con un alto número de clases (en este último caso, hay menos muestras disponibles para cada clase, lo que favorece el sobreajuste).

Datos	N_c	D	P. Esf.	P. Diag.	P. Comp.	Scott	KNN	SVM
Pima	2	8	71.22	71.74	75.13	73.31	73.18	76.47
Wine	3	13	75.84	98.31	99.44	99.44	76.97	100
Landsat	6	36	89.45	90.10	86.10	84.95	90.60	90.90
Optdigits	10	64	97.89	93.82	93.54	93.54	94.38	97.16
Letter	26	16	95.23	91.93	92.77	94.90	95.20	97.55

Cuadro 4.1: Tasa de acierto de la clasificación de Parzen sobre bases de datos públicas. N_c es el número de clases y D la dimensión de los datos.

Para tratar de paliar las limitaciones de estos modelos (la falta de flexibilidad de los modelos esférico y diagonal, y el riesgo de sobreajuste del diagonal y del completo) se propone, a modo de alternativa, un camino intermedio que evita las contrapartidas de ambas opciones. Esta nueva versión se basa en el preprocesado PCA de las muestras de cada clase, para hacerlas adecuadas para un modelo esférico. Se hace uso de la propiedad de las funciones densidad de probabilidad según la cual se cumple $p(\mathbf{A}^T\mathbf{x}) = |\mathbf{A}|^{-1}p(\mathbf{x})$. Sean las matrices de blanqueado de cada clase denotadas por $\mathbf{B}_l$, de modo que $\mathbf{B}_l^T\mathbf{x}$ tiene sus componentes normalizadas y decorrelacionadas para esa clase. El algoritmo de clasificación de acuerdo con este procedimiento viene descrito en la Figura 4.1.

Obtener las matrices de blanqueado $\{\mathbf{B}_l\}$, $l = 1, \ldots, N_c$
Para cada $\mathbf{x}$ del conjunto de test
    Proyectar la muestra $\mathbf{x}$, $\tilde{\mathbf{x}} = \mathbf{B}_l^T\mathbf{x}$
    Evaluar los modelos optimizado mediante ML-LOO $\hat{p}(\tilde{\mathbf{x}}|c_l)$
    Estimar las probabilidades $\hat{p}(\mathbf{x}|c_l) = |\mathbf{B}_l|\hat{p}(\tilde{\mathbf{x}}|c_l)$
    Decidir la clase de la forma: $\hat{y} = \text{máx}_l\, \hat{p}(\mathbf{x}|c_l)$
Fin

Figura 4.1: Algoritmo híbrido para clasificación de Parzen con PCA.

Existe un problema en la aplicación de este método cuando la matriz de covarianza empírica de los datos de una de las clases es singular (esto ocurre, por ejemplo, siempre que haya más dimensiones que muestras). En tal caso, cada matriz de blanqueado $\mathbf{B}_l$ no es cuadrada, pero tiene tantas filas como autovalores no negativos tiene la matriz de covarianza empírica de la clase c_l. Esto es un problema, ya que en el tercer paso del algoritmo (Fig. 4.1) se calcula el determinante de dicha matriz. Por ello proponemos tomar la siguiente estimación del determinante cuando la matriz no es cuadrada

$$|\mathbf{B}| \approx \sqrt{|\mathbf{B}^T\mathbf{B}|}$$

Los nuevos resultados se muestran en la Tabla 4.2, junto con los que se obtuvieron para los modelos esférico, diagonal y completo. Como puede apreciarse, las prestaciones del modelo híbrido se acercan, en general, al mejor de los resultados de entre los otros tres.

Datos	Esférico	Diagonal	Completo	Híbrido
Pima	71.22	71.74	75.13	73.05
Wine	75.84	98.31	99.44	99.44
Landsat	89.45	90.10	86.10	84.10
Optdigits	97.89	93.82	93.54	96.99
Letter	95.23	91.93	92.77	94.93

Cuadro 4.2: Tasa de acierto de la clasificación de Parzen híbrida comparada con los modelos esférico, diagonal y completo.

Para visualizar el efecto de la elección del ancho de ventana en el resultado de clasificación se ha llevado a cabo el siguiente experimento: se han generado datos en 5 dimensiones, siguiendo una distribución gausiana normalizada, y se han generado las etiquetas de acuerdo con la función $y = \text{sign}(x_4 x_5)$. Usando 1000 muestras para entrenamiento, se ha obtenido el ancho de ventana óptimo a partir de todos los datos, que será usado para modelar ambas clases mediante un modelo esférico. En la Figura 4.2 se muestra el resultado de clasificación sobre conjuntos de test de 500 muestras, en el rango comprendido entre $\hat{\sigma}^2/10$ y $10\hat{\sigma}^2$, siendo $\hat{\sigma^2}$ el ancho obtenido por el criterio ML-LOO. Como puede apreciarse, el mejor resultado de clasificación coincide con el ancho obtenido por el algoritmo ML-LOO, lo que refuerza la idoneidad del criterio de máxima verosimilitud en problemas de clasificación.

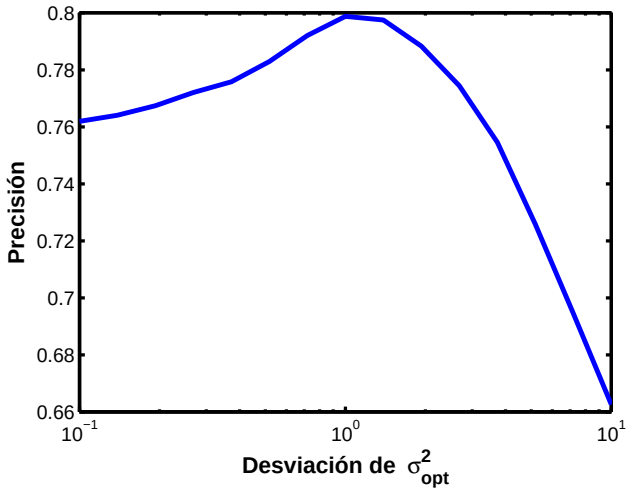

Figura 4.2: Precisión promediada (100 intentos) del clasificador de Parzen en función del ancho de ventana para el problema sintético.

4.2. Aplicación a extracción lineal de características

A continuación se presenta una metodología para extracción de características que plantea el problema de clasificación en términos de teoría de la decisión. En el siguiente apartado se describe la aplicación del criterio de test de hipótesis a la búsqueda de proyecciones lineales óptimas. Posteriormente se relaciona el resultado alcanzado con el método de análisis de discriminantes informativos, ya propuesto en la literatura. Se propondrá igualmente un método que generaliza el resultado para extracción de características mediante modelos GMM, maximizándose la información mutua en lugar del cociente de verosimilitudes. Para finalizar, se presentan algunos resultados de clasificación cuando se usan los métodos propuestos como extractores de características.

4.2.1. Maximización del cociente de verosimilitudes: el algoritmo MaxLT

Un problema de clasificación puede ser enunciado en términos de teoría clásica de la decisión. Inferir la clase a la que pertenece una muestra u observación equivale a la elección de una de entre una serie de hipótesis bajo las cuales se asume que ha sido generada la observación. Esta metodología ha sido ampliamente utilizada en comunicaciones digitales, detección de blancos radar o diagnosis clínica, entre otras aplicaciones [Van Trees, 1968].

En un problema de decisión binaria se ha de elegir entre una de las hipótesis $\mathcal{H}_0$ y $\mathcal{H}_1$ de acuerdo con una determinada observación $\mathbf{x}$. La decisión viene dada por el cociente entre las verosimilitudes de la observación con respecto a las hipótesis, es decir, se decide de acuerdo con el criterio

$$\frac{p(\mathbf{x}|\mathcal{H}_1)}{p(\mathbf{x}|\mathcal{H}_0)} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \lambda \tag{4.7}$$

donde λ es un umbral determinado por algún criterio como el de Neymann-Pearson o el de Bayes [Kay, 1998]. El Lema de Neymann-Pearson establece que dicho test es

óptimo en el sentido de que no existe otro criterio que reduzca simultáneamente los dos tipos de error (la probabilidad de decidir $\mathcal{H}_0$ cuando se da $\mathcal{H}_1$ y viceversa), si las densidades $p(\mathbf{x}|\mathcal{H}_0)$ y $p(\mathbf{x}|\mathcal{H}_1)$ son conocidas [Cover, 2006].

El planteamiento del problema de la extracción de características en términos del test en (4.7) que se propone en este apartado consiste en obtener la transformación lineal $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ tal que haga máximo el test entre $p(\mathbf{z}|\mathcal{H}_1)$ y $p(\mathbf{z}|\mathcal{H}_0)$ para los datos de entrenamiento, es decir

$$LT(\mathbf{z}, y) = \frac{p(\mathbf{z}|\mathcal{H}_1)}{p(\mathbf{z}|\mathcal{H}_0)} \tag{4.8}$$

donde $\mathcal{H}_1$ es la hipótesis de que la muestra $\mathbf{z}$ pertenezca a la clase dada por su etiqueta, y $\mathcal{H}_0$ la hipótesis de que no pertenezca. El objetivo es, por tanto, encontrar la transformación que hace máximo el test en (4.8). El test se lleva a cabo mediante verosimilitudes empíricas, ya que las densidades $p(\mathbf{z}|\mathcal{H}_0)$ y $p(\mathbf{z}|\mathcal{H}_1)$ han de ser estimadas.

A la hora de abordar el caso multiclase, se propone generalizar el test en (4.8) de forma que ahora $\mathcal{H}_0$ abarca todas las clases salvo la correcta (la dada por la etiqueta). Este procedimiento equivale a optar por un esquema de aprendizaje del tipo uno-contra-el-resto, que consiste en construir L clasificadores binarios que discriminan entre cada una de las clases y el resto. Cada uno de estos L clasificadores proporciona un valor positivo si decide la clase l y negativo si decide la complementaria; la decisión final será la procedente de la máquina que mayor valor proporcione a su salida. Este procedimiento ha sido criticado por su carácter heurístico, ya que cada clasificador binario resuelve un problema distinto, y asumir que sus valores de salida son comparables puede constituir una suposición demasiado fuerte [Scholkopf, 2002]. Sin embargo, el procedimiento uno-contra-el-resto nos permite disponer de más muestras para construir los clasificadores que en el caso uno-contra-uno. En nuestro caso, nos permite caracterizar la hipótesis complementaria con la totalidad de las muestras que no pertenecen a esa clase, lo que lo convierte en una alternativa robusta.

Asumiendo las muestras de entrenamiento independientes e idénticamente distribuidas, el test para la totalidad del conjunto de entrenamiento se puede reescribir como

$$LT(\mathbf{Z}, Y) = \frac{\prod_i \hat{p}(\mathbf{z}_i|y_i)}{\prod_i \hat{p}(\mathbf{z}_i|\bar{y}_i)} \tag{4.9}$$

Si nos apoyamos en la monotonicidad del logaritmo, convertimos los productos en sumas. En este caso, pretendemos encontrar $\mathbf{z} = \mathbf{W}^T\mathbf{x}$ que hace máximo

$$\arg\max_{\mathbf{W}} \log LT(\mathbf{Z}, Y) = \arg\max_{\mathbf{W}} \sum_i \left[\log \hat{p}(\mathbf{z}_i|y_i) - \log \hat{p}(\mathbf{z}_i|\bar{y}_i)\right] \tag{4.10}$$

A la hora de estimar las densidades de cada clase, conviene una vez más optar por un modelo de Parzen, optimizado mediante el algoritmo ML-LOO. La hipótesis de pertenencia a la clase c_l viene definida por

$$\hat{p}(\mathbf{z}|c_l) = \frac{1}{N_l} \sum_{i \in I_l} G(\mathbf{z} - \mathbf{z}_i|\sigma_l^2)$$

donde I_l es el conjunto de índices de las muestras pertenecientes a la clase c_l. La hipótesis de no pertenencia viene modelada como

$$\hat{p}(\mathbf{z}|\bar{c}_l) = \sum_{j \neq l} \pi_{j,l} \hat{p}(\mathbf{z}|c_j) \tag{4.11}$$

donde cada $\pi_{j,l}$ es un prior que indica la probabilidad a-priori de pertenecer a c_j sujeto a que no se pertenece a c_l. Este prior puede ser determinado empíricamente a partir del número de muestras de entrenamiento que pertenecen a cada clase, o a partir de cualquier otro tipo de información a priori sobre los datos. En el primer caso, tendríamos: $\pi_{j,l} = \frac{N_j}{N - N_l}$. De esta forma, se puede reescribir (4.11) como

$$\begin{aligned}
\hat{p}(\mathbf{z}|\bar{c}_l) &= \sum_{j \neq l} \frac{N_j}{N - N_l} \hat{p}(\mathbf{z}|c_j) \\
&= \sum_{j \neq l} \frac{N_j}{N - N_l} \frac{1}{N_j} \sum_{i \in I_j} G(\mathbf{z} - \mathbf{z}_i|\sigma_{c_j}^2) \\
&= \frac{1}{N - N_l} \sum_{j \neq l} \sum_{i \in I_j} G(\mathbf{z} - \mathbf{z}_i|\sigma_{c_j}^2)
\end{aligned}$$

El cociente de verosimilitudes queda entonces de la forma

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \sum_{i=1}^{N} \left[\log \frac{1}{N_l} \sum_{j \in I_l} G(\mathbf{z}_i - \mathbf{z}_j | \sigma_l^2) - \log \frac{1}{N - N_l} \sum_{k \neq l} \sum_{j \in I_k} G(\mathbf{z}_i - \mathbf{z}_j | \sigma_k^2) \right]$$

Disponemos de todos los elementos necesarios para maximizar el cociente. Supuesto un esquema de extracción lineal, con $\mathbf{z} = \mathbf{W}^T \mathbf{x}$, el gradiente respecto a la transformación viene dado por

$$\nabla_{\mathbf{W}} \log LT(\mathbf{Z}, Y) = \sum_i \nabla_{\mathbf{W}} \left[\log \hat{p}(\mathbf{z}_i | y_i) - \log \hat{p}(\mathbf{z}_i | \bar{y_i}) \right]$$

Como puede verse, cada muestra influye en cada una de las otras a través de sólo uno de los miembros, ya que los modelos de las densidades de ambas hipótesis están construidos con conjuntos de muestras disjuntos. Desarrollando, llegamos a

$$\begin{aligned} \nabla_{\mathbf{W}} \log LT(\mathbf{Z}, Y) = \sum_i \Bigg[& \frac{1}{\hat{p}(\mathbf{z}_i | y_i)} \frac{1}{N_{y_i}} \sum_{j \in I_{y_i}} \nabla_{\mathbf{W}} G(\mathbf{z} - \mathbf{z}_i | \sigma_{y_i}^2) \\ & - \frac{1}{\hat{p}(\mathbf{z}_i | \bar{c}_{y_i})} \frac{1}{N - N_{y_i}} \sum_{c_j \neq y_i} \nabla_{\mathbf{W}} G(\mathbf{z} - \mathbf{z}_i | \sigma_j^2) \Bigg] \end{aligned} \tag{4.12}$$

De esta forma, ya contamos con la expresión de la derivada que nos permitirá construir un algoritmo de ascenso por gradiente para maximizar el cociente de verosimilitudes en (4.9). Este ascenso por gradiente está acompañada de una ortogonalización de Gram-Schmidt como la que se aplicó al método MMI en el Capítulo 2.

4.2.2. Relación con análisis de discriminantes informativos

El análisis de discriminantes informativos (*Informative Discriminant Analysis*, IDA) ha sido propuesto para extracción de características haciendo uso de modelos de Parzen [Kaski, 2003], [Peltonen, 2005]. Este método busca la función de transformación $\mathbf{z} = \mathbf{W}^T \mathbf{x}$ que maximiza la verosimilitud

$$\log L(\mathbf{Z}, Y) = \sum_i \log \hat{p}(y_i | \mathbf{z}_i) \tag{4.13}$$

La densidad condicional se estima como

$$\hat{p}(y_i|\mathbf{z}_i) = \frac{\hat{p}(\mathbf{z}_i|c_{y_i})}{\sum_l \hat{p}(\mathbf{z}_i|c_l)} \tag{4.14}$$

donde las $\hat{p}(\mathbf{z}_i|c_l)$ son modelos de Parzen. El método de extracción de características consiste en encontrar la matriz de transformación $\mathbf{W}$ que maximiza la verosimilitud en (4.13). Este criterio puede ser manipulado de la forma

$$\begin{aligned}
\hat{\mathbf{W}} &= \arg\max_{\mathbf{W}} \sum_i \log \frac{\hat{p}(\mathbf{z}_i|y_i)}{\sum_l \hat{p}(\mathbf{z}_i|c_l)} = \arg\max_{\mathbf{W}} \sum_i \log \frac{1}{1 + \frac{\sum_{c_l \neq y_i} \hat{p}(\mathbf{z}_i|c_l)}{\hat{p}(\mathbf{z}_i|y_i)}} \\
&= \arg\min_{\mathbf{W}} \sum_i \log\big[1 + \frac{\sum_{c_l \neq y_i} \hat{p}(\mathbf{z}_i|c_l)}{\hat{p}(\mathbf{z}_i|y_i)}\big] \\
&= \arg\min_{\mathbf{W}} \sum_i \log \frac{\sum_{c_l \neq y_i} \hat{p}(\mathbf{z}_i|c_l)}{\hat{p}(\mathbf{z}_i|y_i)} \\
&= \arg\max_{\mathbf{W}} \sum_i \big[\log \hat{p}(\mathbf{z}_i|y_i) - \log \sum_{c_l \neq y_i} \hat{p}(\mathbf{z}_i|c_l)\big]
\end{aligned} \tag{4.15}$$

Si desarrollamos cada $\hat{p}(\mathbf{z}_i|c_l)$ llegamos a

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \sum_i \left[\log \frac{1}{N_l} \sum_{j \in I_l} G(\mathbf{z}_i - \mathbf{z}_j|\sigma_l^2) - \log \sum_{k \neq l} \frac{1}{N_k} \sum_{j \in I_k} G(\mathbf{z}_i - \mathbf{z}_j|\sigma_k^2)\right] \tag{4.16}$$

Podemos ahora llevar a cabo una comparación entre los criterios (4.15) y (4.2). Como puede apreciarse, la única diferencia radica en el peso que adopta la hipótesis complementaria en el funcional. En el caso de (4.2), el argumento de cada logaritmo en el segundo término es el correspondiente a la probabilidad de pertenencia a la clase complementaria. En (4.15), sin embargo, el argumento consiste en la suma de las probabilidades de pertenencia a cada una de las clases. Esto se traduce en que el método IDA asigna una mayor penalización a las situaciones de no pertenencia, debido a la propiedad dada por la cota de la unión: $\sum_n p(a_n) \geq p(\sum_n a_n)$.

4.2.3. Extensión a modelos semiparamétricos

En este apartado se plantea un esquema de extracción de características que hace uso de modelos de FPD de tipo GMM, y que trata de maximizar la información

mutua entre las proyecciones y las clases, en lugar del test de verosimilitud entre hipótesis. La información mutua entre las componentes y las clases se puede expresar en términos de entropía como $I(\mathbf{z}, y) = h(\mathbf{z}) - h(\mathbf{z}|y)$; la estimación de la entropía $\hat{h}(\mathbf{W}^T\mathbf{x})$ a partir de un modelo GMM es posible de forma muestreada, análogamente a la aproximación que se mostró en (3.11) para modelos de Parzen. A partir de esta estimación, es posible por tanto la aproximación de la información mutua y la maximización de la misma, ya que dicha aproximación es derivable con respecto a la transformación.

Un modelo GMM viene dado por la expresión

$$\hat{p}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k G(\mathbf{x}|\boldsymbol{\theta}_k) \tag{4.17}$$

Un esquema de maximización de la información mutua basado en este tipo de modelos GMM tiene el inconveniente de que conforme los datos son proyectados de acuerdo a $\mathbf{W}_n$, siendo n la enésima iteración del ascenso por gradiente, el modelo de mezcla de máxima verosimilitud debería ser recalculado en cada paso. Esto supondría disparar la complejidad computacional del problema. En lugar de ello, se va a suponer que los parámetros del modelo GMM para $\mathbf{x}$ sigue siendo válido para $\mathbf{z}$, una vez proyectados aquellos, es decir $\mathbf{m}'_k = \mathbf{W}^T\mathbf{m}_k$, $\mathbf{C}'_k = \mathbf{W}^T\mathbf{C}_k\mathbf{W}$, siendo $\boldsymbol{\theta}_k = \{\mathbf{m}_k, \mathbf{C}_k\}$ los parámetros de la k-ésima gausiana en el espacio de $\mathbf{x}$ y $\boldsymbol{\theta'}_k = \{\mathbf{m'}_k, \mathbf{C'}_k\}$ los del espacio de $\mathbf{z}$

La estimación de la entropía, para un modelo de mezcla con K gausianas, tendría la forma

$$\hat{h}(\mathbf{z}) = -\frac{1}{N}\sum_{i=1}^{N} \log \sum_{k=1}^{K} \alpha_k G(\mathbf{z}_i|\boldsymbol{\theta'}_k) \tag{4.18}$$

El gradiente de esta entropía respecto de la matriz $\mathbf{W}$ viene dada por

$$\nabla_{\mathbf{W}}\hat{h}(\mathbf{W}^T\mathbf{x}) = -\frac{1}{N}\sum_{i} \frac{1}{\hat{p}(\mathbf{z}_i)} \sum_{k} \alpha_k \nabla_{\mathbf{W}} G(\mathbf{z}_i|\boldsymbol{\theta'}_k) \tag{4.19}$$

donde el gradiente de cada gausiana tiene el valor

$$\begin{aligned}\nabla_{\mathbf{W}} G(\mathbf{z}_i|\boldsymbol{\theta}_k) = G(\mathbf{z}_i|\boldsymbol{\theta'}_k)\big[\big(\mathbf{C'}_k^{-1}(\mathbf{z}_i - \mathbf{m'}_k)(\mathbf{z}_i - \mathbf{m'}_k)^T - \mathbf{I}\big)\mathbf{C'}_k^{-1}\mathbf{W}^T\mathbf{C}_k \\ - \mathbf{C'}_k^{-1}(\mathbf{z}_i - \mathbf{m'}_k)(\mathbf{x}_i - \mathbf{m}_k)^T\big]\end{aligned} \tag{4.20}$$

El algoritmo, que denominaremos MMI-GMM, funciona por tanto mediante optimización con ascenso por gradiente, donde la derivada de la información mutua viene dada por

$$\nabla_{\mathbf{W}} I(\mathbf{z}|y) = \nabla_{\mathbf{W}} \hat{h}(\mathbf{z}) - \sum_{l=1}^{N_c} P(c_l) \nabla_{\mathbf{W}} \hat{h}(\mathbf{z}|c_l) \tag{4.21}$$

En realidad, en 4.21 pueden emplearse estimadores de la entropía basados en ventanas de Parzen en lugar de modelos GMM. Sin embargo, esto lleva acarreado un gran coste computacional: la complejidad de la evaluación de la entropía en el primer término de (4.21) es $O(N^2)$, mientras que en el caso de ser estimada con GMM es $O(ND)$. Esta diferencia es sustancial, ya que supone complejidad cuadrática frente a complejidad lineal con el número de muestras. El método MaxLT hace uso de estimadores de la verosimilitud de complejidad idéntica a los estimadores de entropía propuestos. Sin embargo, la diferencia fundamental entre MaxLT y MMI-GMM es el hecho de que en el primer caso no tenemos ningún término en (4.12) que tenga complejidad $O(N^2)$. Dicho de otro modo, no tenemos ningún término que suponga utilizar la totalidad de las muestras para estimar la probabilidad (o verosimilitud) en cada una de las mismas.

4.2.4. Experimentos

En este apartado se analizan las prestaciones de los métodos de extracción de características presentados sobre bases de datos públicas. A continuación se lleva a cabo una comparación entre los métodos MaxLT e IDA. Después, se muestran algunos resultados de aplicar el método MMI-GMM.

Prestaciones de MaxLT e IDA

El algoritmo que se ha implementado tanto para la maximización del test de hipótesis (MaxLT) o la verosimilitud de los datos respecto a las etiquetas (IDA) opera en dos fases, de la siguiente forma:

1. En primer lugar, y dado que no es posible conocer el ancho óptimo de las ventanas de Parzen para el vector de salida $\mathbf{z}$, éste es fijado a priori de forma heurística. Un criterio razonable es utilizar el ancho $\hat{\sigma}^2 = \frac{d_{max}^2}{4}$, siendo d_{max} la distancia máxima entre cualquier par de puntos del conjunto de entrenamiento. De esta forma garantizamos que existirá una interacción entre todos los pares de puntos, por lo que todas las evaluaciones del tipo $G(\mathbf{z}_i - \mathbf{z}_j|\sigma_l^2)$ tendrán un valor no nulo. En este primer paso se impone el mismo ancho para todas las clases.

2. Una vez la primera fase ha convergido a su valor máximo, utilizando un ancho de ventana no óptimo, se asume que la proyección encontrada está cerca de la óptima, por lo que es razonable calcular el ancho óptimo, mediante el método ML-LOO, a partir de los datos proyectados. Una vez se ha obtenido el ancho óptimo, se deja evolucionar de nuevo el algoritmo, a partir de la proyección obtenida en la fase 1, hasta su nueva convergencia.

Esta metodología supone una simplificación de la técnica de temple determinista (*Deterministic Annealing*) [Rose, 1999] que ha sido usada con éxito en algoritmos similares [Torkkola, 2003]. De esta forma, se evita el procedimiento de validación cruzada propuesta por los autores del método IDA para la obtención de $\hat{\sigma}^2$.

A las bases de datos se les ha aplicado una decorrelación previa con PCA, de forma que se obtenga una esfericidad en los datos que haga apropiado el uso de modelos esféricos. Además, esto hace que las proyecciones de salida estén también decorrelacionadas y normalizadas, lo que hace más fácil la búsqueda de hiperparámetros cuando se usa una SVM.

Clasificador	1	2	3	4	5	6
MaxLT/KNN	17.50	36.70	50.00	63.72	70.80	77.25
MaxLT/SVM (1 vs 1)	18.07	41.37	52.57	**67.50**	73.90	78.45
MaxLT/SVM (1 vs rest)	5.77	28.57	46.45	66.87	**74.20**	80.35
MaxLT/Parzen	18.45	**42.80**	**55.25**	67.30	73.32	79.37
IDA/KNN	18.05	36.55	49.37	63.12	70.75	76.85
IDA/SVM (1 vs 1)	18.15	40.17	53.97	67.20	74.05	**80.67**
IDA/SVM (1 vs rest)	7.82	33.77	50.32	66.65	73.95	80.27
IDA/Parzen	**18.70**	42.67	55.22	67.37	73.32	79.37

Cuadro 4.3: Resultados de clasificación para la base de datos Letter tras usar MaxLT e IDA.

Clasificador	1	2	3	4	5	6
MaxLT/KNN	41.50	67.15	80.40	83.40	83.70	82.45
MaxLT/SVM (1 vs 1)	50.35	75.90	**83.75**	85.95	**87.95**	**86.80**
MaxLT/SVM (1 vs rest)	42.80	**76.05**	83.25	**86.15**	87.75	85.10
MaxLT/Parzen	47.30	73.85	82.10	85.10	87.65	84.65
IDA/KNN	39.80	65.00	79.35	82.05	84.05	84.15
IDA/SVM (1 vs 1)	**51.15**	73.50	83.15	84.85	86.50	86.55
IDA/SVM (1 vs rest)	39.60	73.55	83.25	84.30	84.60	85.25
IDA/Parzen	48.95	71.85	82.80	85.30	85.55	86.05

Cuadro 4.4: Resultados de clasificación para la base de datos Landsat tras usar MaxLT e IDA.

Clasificador	1	2	3	4	5	6
MaxLT/KNN	20.91	41.57	57.34	63.44	64.66	74.98
MaxLT/SVM (1 vs 1)	27.45	49.97	**66.77**	72.03	73.32	80.18
MaxLT/SVM (1 vs rest)	11.42	42.40	64.34	71.65	70.75	80.37
MaxLT/Parzen	28.03	50.93	66.20	71.33	69.21	79.28
IDA/KNN	20.78	42.66	56.70	63.50	68.12	76.40
IDA/SVM (1 vs 1)	28.09	51.31	65.68	**72.26**	**77.23**	**80.56**
IDA/SVM (1 vs rest)	10.26	40.09	60.04	71.46	76.40	80.37
IDA/Parzen	**28.16**	**51.38**	65.30	71.20	75.11	79.15

Cuadro 4.5: Resultados de clasificación para la base de datos Isolet tras usar MaxLT e IDA.

Clasificador	1	2	3	4	5	6
MaxLT/KNN	32.00	49.97	74.90	85.09	89.15	92.38
MaxLT/SVM (1 vs 1)	39.57	57.27	**80.08**	**88.09**	89.32	88.70
MaxLT/SVM (1 vs rest)	32.50	57.15	79.19	87.59	89.54	92.93
MaxLT/Parzen	**39.29**	**58.15**	79.74	87.76	**90.76**	93.04
IDA/KNN	26.54	50.70	76.35	84.92	88.81	91.99
IDA/SVM (1 vs 1)	33.00	56.43	79.91	88.04	87.09	92.60
IDA/SVM (1 vs rest)	29.38	56.04	79.08	87.76	89.37	92.15
IDA/Parzen	33.89	56.76	79.47	87.59	90.37	**93.10**

Cuadro 4.6: Resultados de clasificación para la base de datos Optdigits tras usar MaxLT e IDA.

En las Tablas 4.3 hasta 4.6 se muestran los resultados de ambos métodos para varias bases de datos públicas. Se han evaluado las prestaciones en clasificación mediante los métodos uno-contra-uno (tanto con KNN como con SVM), uno-contra-todos también con SVM, y clasificación de Parzen con ventana esférica, optimizada de acuerdo con el método ML-LOO. De nuevo, los hiperparámetros de la SVM han sido ajustados mediante un procedimiento de validación cruzada con tres subconjuntos de los datos de entrenamiento. Los resultados sobre la base de datos Letter son más pobres que los mostrados en el Capítulo 2: esto se debe a que por razones de complejidad computacional el tamaño muestral de entrenamiento ha sido reducido a la mitad, perdiéndose información discriminante.

Los resultados muestran unas prestaciones ligeramente superiores por parte de MaxLT respecto a los proporcionados por IDA, aunque esta diferencia es mínima por lo que pueden considerarse métodos de potencia equivalente.

Los resultados de clasificación ponen de relieve la inferioridad del método de clasificación uno-contra-el-resto llevado a cabo, en este caso, mediante SVM. Aun así, queda de manifiesto el hecho de que la extracción de características presentada, guiada por el criterio de mejor clasificación uno-contra-el-resto, proporciona un conjunto de proyecciones que da buenos resultados de clasificación aun cuando se usa un criterio que minimiza el error uno-contra-uno. Al igual que en el Apartado 4.1.2, donde se mostraron los resultados de clasificación de Parzen en comparación con los obtenidos por otros métodos, aquí hay que volver a destacar los resultados competitivos obtenidos por dicho clasificador, especialmente en las bases de datos Letter y Optdigits. Dichos resultados pueden ser interpretados en el sentido de que el método MaxLT se basa en la maximización del test de verosimilitudes, que es de hecho el criterio evaluado en clasificación de Parzen (aunque en MaxLT se maximice el test uno-contra-el-resto). Es por tanto razonable que, habiéndose diseñado apropiadamente el ancho de los modelos de Parzen, éstos tengan una buena capacidad de generalización. De esta forma, la transformación que hace máximo el test para los datos de entrenamiento sigue siendo válida para los datos de test.

Prestaciones de MMI-GMM

En las Tablas 4.7 hasta 4.10 se muestran los resultados de clasificación cuando se lleva a cabo la extracción de características mediante la maximización de la información mutua con GMMs presentada en el Apartado 4.2.3. Se han evaluado las prestaciones del método para distintos grados de reducción de la dimensión, así como para distintos grados de complejidad de los modelos GMM usados. El número de gausianas usadas para las mezclas va desde una hasta cuatro, como se muestra en las tablas. De nuevo se proporcionan los resultados tanto de KNN ($K = 1$) como de SVM. En realidad, el procedimiento seguido en el conjunto de experimentos es subóptimo, debido al hecho de la l-ésima fila de cada tabla muestra el resultado de aplicar una mezcla de l gausianas a cada subconjunto de muestras correspondiente a cada una de las clases. La misma complejidad del modelo es usada para modelar la densidad de la totalidad de los datos, necesaria para estimar la $\hat{h}(\mathbf{x})$ de la estimación de la información mutua en (4.21). Cada subconjunto de muestras será susceptible de ser modelado con mezclas de distinto orden de complejidad. Elegir, por tanto, el orden del modelo óptimo para cada subconjunto aumenta considerablemente la complejidad del problema, y exige la aplicación de algún criterio de penalización de la complejidad de forma independiente al subconjunto de cada clase. Aun así se obtienen resultados que son, en general, mejores que los mostrados para MaxLT e IDA, que usan modelos no paramétricos de FDP. Hay dos razones por las que puede haberse alcanzado este resultado. La primera es el hecho de que GMM puede proporcionar un modelo más preciso que un modelo de Parzen esférico. La segunda es que en el método MaxLT la relación entre la entrada y la salida del extractor de características está caracterizado por la relación $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, mientras que en el método MMI-GMM se tiene en cuenta, además, la relación entre los parámetros del modelo a la entrada y a la salida del extractor, de forma que $\mathbf{m}'_k = \mathbf{W}^T\mathbf{m}_k$, $\mathbf{C}'_k = \mathbf{W}^T\mathbf{C}_k\mathbf{W}$.

Nº de mezclas	1	2	3	4	5	6
1 (KNN)	17.82	35.17	44.30	60.67	71.20	78.97
1 (SVM)	18.02	37.00	48.72	64.35	74.10	80.10
2 (KNN)	17.50	37.60	55.27	70.70	76.62	82.80
2 (SVM)	**19.37**	41.10	**58.65**	**73.17**	77.85	84.60
3 (KNN)	18.00	39.07	55.85	68.07	76.25	83.75
3 (SVM)	19.27	**41.95**	58.27	66.40	78.45	**86.17**
4 (KNN)	17.45	39.45	54.10	69.77	78.32	83.72
4 (SVM)	19.12	40.95	57.90	71.95	**79.70**	85.20

Cuadro 4.7: Precisión obtenida para la base de datos Letter tras usar MMI-GMM.

Nº de mezclas	1	2	3	4	5	6
1 (KNN)	56.60	66.25	83.20	83.30	85.40	83.15
1 (SVM)	**67.00**	72.75	**86.35**	**87.75**	**87.80**	87.40
2 (KNN)	58.10	79.75	83.50	83.35	82.15	81.60
2 (SVM)	64.40	**83.80**	85.35	86.15	86.60	87.25
3 (KNN)	57.65	78.70	84.05	84.95	82.25	81.80
3 (SVM)	64.45	83.50	85.95	87.05	86.75	87.25
4 (KNN)	56.50	79.65	82.85	84.85	83.20	82.15
4 (SVM)	64.45	83.25	86.30	87.25	86.35	**87.60**

Cuadro 4.8: Precisión obtenida para la base de datos Landsat tras usar MMI-GMM.

Nº de mezclas	1	2	3	4	5	6
1 (KNN)	21.55	39.77	49.01	60.36	67.99	70.30
1 (SVM)	29.76	49.58	60.36	68.63	74.34	74.92
2 (KNN)	23.41	44.13	59.14	65.94	71.97	75.37
2 (SVM)	**31.82**	56.83	67.80	74.34	79.22	80.56
3 (KNN)	19.76	47.98	63.69	72.67	75.95	80.24
3 (SVM)	30.92	**59.46**	71.78	**80.50**	83.90	79.86
4 (KNN)	22.19	48.36	65.43	69.98	80.44	82.42
4 (SVM)	30.08	58.24	**72.23**	77.23	**84.48**	**87.11**

Cuadro 4.9: Precisión obtenida para la base de datos Isolet tras usar MMI-GMM.

Nº de mezclas	1	2	3	4	5	6
1 (KNN)	33.33	56.54	72.79	85.20	90.32	90.48
1 (SVM)	41.51	62.94	78.63	**88.37**	89.48	85.03
2 (KNN)	34.00	65.00	76.35	84.59	88.65	91.49
2 (SVM)	43.13	71.23	81.41	87.59	90.09	85.87
3 (KNN)	31.50	65.33	82.69	84.42	89.48	92.49
3 (SVM)	42.07	**74.40**	**85.64**	87.37	**91.65**	**93.77**
4 (KNN)	35.62	63.16	82.14	85.25	88.43	90.93
4 (SVM)	**43.24**	71.23	**85.64**	76.68	88.15	87.87

Cuadro 4.10: Precisión obtenida para la base de datos Optdigits con MMI-GMM.

4.3. Aplicación a análisis de componentes independientes

Como se ha descrito en el Apartado 1.3.4, ICA busca una matriz cuadrada $\mathbf{W}_{ICA}$ que, aplicada a una variable multidimensional $\mathbf{x}$ de la forma $\mathbf{z} = \mathbf{W}_{ICA}^T\mathbf{x}$, proporciona en $\mathbf{z}$ un conjunto de señales estadísticamente independientes. En este apartado se presenta un método ICA que hace uso de la estimación de FDP mediante ventanas de Parzen optimizadas como se estudió en el Capítulo 3. A continuación se describe la función de contraste con la que trabaja el método propuesto. Tras esto se describirá el método de búsqueda con restricciones de ortonormalidad. Finalmente se presentan algunos resultados experimentales.

4.3.1. Función de coste a minimizar

Una matriz ICA puede ser obtenida mediante la aplicación del criterio de mínima información mutua entre las señales resultantes

$$\mathbf{W}_{ICA} = \arg\min_{\mathbf{W}} I(\mathbf{z}) = \arg\min_{\mathbf{W}} \left[\sum_{k=1}^{D} h(z_k) - h(\mathbf{z})\right] \tag{4.22}$$

Es una práctica común forzar que la matriz $\mathbf{W}$ sea ortonormal. De esta forma se reduce el ámbito de búsqueda de la matriz, y da lugar a un mejor condicionamiento del problema. En ese caso, tenemos

$$h(\mathbf{z}) = h(\mathbf{x}) + \log|\det \mathbf{W}| = h(\mathbf{x})$$

dado que $\det \mathbf{W} = 1$. Como $h(\mathbf{x})$ es constante respecto de $\mathbf{W}$, podemos reexpresar el criterio en (1.26) de la forma

$$\mathbf{W}_{ICA} = \arg\min_{\mathbf{W}} \sum_k h(z_k) = \arg\min_{\{\mathbf{w}_k\}} \sum_k h(z_k)$$

Cada entropía $h(z_k)$ puede ser estimada de acuerdo con el procedimiento LOO que se describió en el Apartado 3.3. Concretamente, se hará uso del modelo de Parzen

con matriz de covarianza completa, en el que no hay restricciones sobre la forma de ésta. Para su aplicación a ICA, necesitamos establecer el conjunto de gradientes $\nabla_{\mathbf{w}_k}\hat{h}_{-1}(z_k) = \dfrac{d\hat{h}_{-1}(z_k)}{d\mathbf{w}_k}$. La derivada de la entropía viene dada por

$$\nabla_{\mathbf{w}_k}\hat{h}_{-1}(z_k) = \frac{1}{N(N-1)}\sum_i \frac{1}{\hat{p}(z_{k,i})}\sum_{j\neq i} G(z_{k,i}-z_{k,j}|\sigma_k^2)\times \\ \times \Big[\frac{1}{\sigma_k^2}\Big(\frac{1}{\sigma_k^2}(z_{k,i}-z_{k,j})(z_{k,i}-z_{k,j})^T - 1\Big)\mathbf{w}_k\mathbf{C}_x - \frac{1}{\sigma_k^2}(z_{k,i}-z_{k,j})(\mathbf{x}_i-\mathbf{x}_j)^T\Big] \tag{4.23}$$

donde $z_{k,i}$ es la componente k-ésima de la muestra $\mathbf{z}_i$. Para llegar a (4.23) se ha asumido una relación entre la matriz de covarianza de la ventana a la entrada y a la salida de la forma $\sigma_k^2 = \mathbf{w}_k^T\mathbf{C}_\mathbf{x}\mathbf{w}_k$.

Dado que $\mathbf{W} = \{\mathbf{w}_k\}$, $k = 1,\ldots,D$, el gradiente respecto a $\mathbf{W}$ es una matriz compuesta por los D gradientes calculados en (4.23).

La principal ventaja del método propuesto es que no exige la decorrelación previa de los datos. Muchos de los métodos ICA propuestos en la literatura plantean el procedimiento ICA como dos transformaciones lineales sucesivas. La primera de ellas aplica PCA, dando lugar a un conjunto de variables decorrelacionadas. La segunda de ellas consiste en una matriz de rotación ortonormal que, además de decorrelación, consigue independencia estadística para las variables. El método propuesto en este apartado no necesita esta restricción previa de correlación, por lo que puede aplicarse directamente sobre los datos.

4.3.2. Reparametrización mediante álgebra de Lie

Cuando se aborda un problema de optimización en el que el argumento del funcional es una matriz cuadrada, y se imponen restricciones del tipo $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, resulta muy útil usar la parametrización mediante álgebra de Lie. Los métodos ICA constan, generalmente, de dos pasos que tienen lugar secuencialmente. El primer paso decorrelaciona las señales: es el proceso que conocemos como "blanqueado". En el segundo, se aplica una matriz de rotación que consigue independencia estadística además de

decorrelación. Una matriz así ha de cumplir la restricción $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ para que se mantenga la decorrelación.

Habitualmente, la rotación en ICA se hace de forma que minimice un determinado funcional $J = J(\mathbf{W})$. El descenso por gradiente es la técnica más habitual para su optimización, lo que implica aplicar una actualización del tipo: $\mathbf{W}_{n+1} = \mathbf{W}_n + \eta\nabla_{\mathbf{W}}J$. Sin embargo, la matriz $\mathbf{W}_{n+1}$ puede dejar de cumplir la restricción, por lo que se debe aplicar una ortogonalización de Gram-Schmidt posterior, o bien introducir un término en el funcional J que penalice la no-ortogonalidad.

El álgebra de Lie proporciona una alternativa elegante a este problema, ya que nos muestra un mecanismo para que el gradiente pueda moverse por la superficie geodésica correspondiente al grupo algebraico de las matrices ortonormales [Plumbley, 2004].

Una matriz ortonormal de rotación tiene la forma

$$\mathbf{\Omega} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{4.24}$$

Y se cumple que el producto de matrices de rotación es también una matriz de rotación

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} = \begin{pmatrix} \cos(\theta+\phi) & \sin(\theta+\phi) \\ -\sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix} \tag{4.25}$$

El operador exponencial matricial nos permite expresar la matriz en (4.24) como

$$\mathbf{\Omega} = \exp\mathbf{\Theta} = \exp\begin{pmatrix} 0 & \theta \\ -\theta & 0 \end{pmatrix} \tag{4.26}$$

Para clarificar la notación, llamaremos dominio Ω al de las matrices ortonormales, y Θ al dominio transformado por el operador logaritmo matricial, de forma que $\mathbf{\Omega} = \exp(\mathbf{\Theta})$.

Como puede verse, no queda más que reparametrizar el funcional a minimizar para que esté en función del dominio Θ. En este dominio, la suma que implica el

ascenso por gradiente se traduce, en el dominio Ω, en multiplicación de matrices de rotación, por lo que no se pierde en ningún momento la restricción de ortonormalidad.

El algoritmo de descenso por gradiente tiene ahora la forma que se muestra en la Tabla 4.3.

$n = 0$
$\mathbf{\Theta}_0 = \mathbf{0}$
$\mathbf{\Omega}_0 = \mathbf{I}$
Repetir
    Aplicar el paso del gradiente: $\mathbf{\Theta}_{n+1} = \mathbf{\Theta}_n + \eta \nabla_{\mathbf{\Theta}} J$;
    Pasar al dominio Ω mediante $\mathbf{\Omega}_{n+1} = \exp(\mathbf{\Theta}_{n+1})$;
    Desplazar en el dominio $\mathbf{\Omega}$ mediante $\mathbf{W}_{n+1} = \mathbf{\Omega}\mathbf{W}_n$;
    n = n+1;
Fin

Figura 4.3: Algoritmo para descenso de gradiente en grupos de Lie.

El gradiente del funcional respecto a los elemento del dominio Θ viene dado por [Plumbley, 2004]

$$\nabla_\Theta J = (\nabla_{\mathbf{\Omega}} J)\mathbf{\Omega}^T - \mathbf{\Omega}(\nabla_\Omega J)^T \tag{4.27}$$

4.3.3. Pruebas y resultados

Para evaluar el correcto funcionamiento del método propuesto, que llamaremos ParzenICA, se ha generado un conjunto de señales sintéticas que el algoritmo se encargará de separar. Las señales sintéticas se muestran en la Figura 4.4, y constan de una sinusoide, una señal cuadrada, una señal diente de sierra y un ruido gausiano. El resultado de mezclar las señales mediante una matriz aleatoria se muestra en la Figura 4.5. Por último, en la Figura 4.6 se muestra el resultado de aplicar el algoritmo. Como puede apreciarse, las señales han sido apropiadamente separadas, lo que demuestra la fiabilidad del algoritmo.

Para analizar la evaluación de las entropías involucradas, en la Figura 4.7 se muestra la evolución de la entropía de cada proyección conforme transcurre la resolución de la separación ciega.

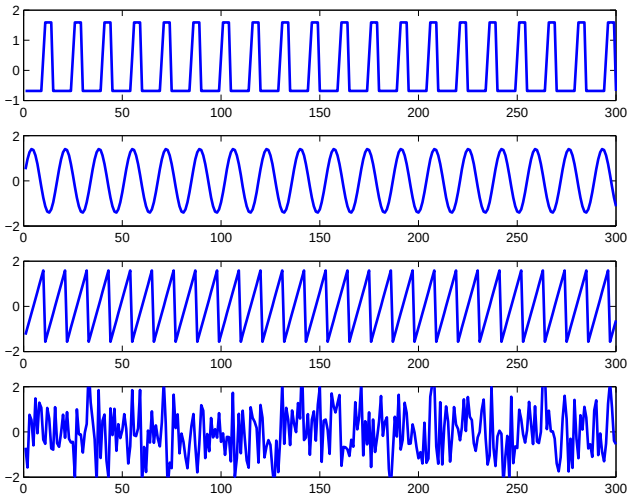

Figura 4.4: Señales sintéticas empleadas para el experimento.

Para probar la eficacia del algoritmo con datos reales, se ha aplicado ParzenICA a la separación de señales de electrocardiograma del abdomen y tórax de una mujer embarazada [Moor, 1997]. Estos datos son frecuentemente usados para probar la eficacia de un método ICA, ya que debe ser capaz de separar las señales correspondientes a la actividad cardíaca de la madre, a la del bebé, y las señales ruidosas que aparecen en el proceso de adquisición [Kraskov, 2004]. Los datos consisten en ocho lecturas de otros tantos electrodos, muestreados a 500 Hz durante 5 segundos. Además, se compara el resultado del algoritmo propuesto con el proporcionado por el método FastICA, que es un método de referencia. Las señales originales pueden verse en la Figura 4.8. El resultado de la separación del algoritmo FastICA es mostrado en la Figura 4.9 a modo de comparación. Los resultados de ParzenICA se muestran en la Figura 4.10.

En la Figura 4.10 se aprecia que las señales correspondientes a los pulsos cardíacos de madre (dos primeras señales) y feto (quinta señal) han sido obtenidas con un nivel

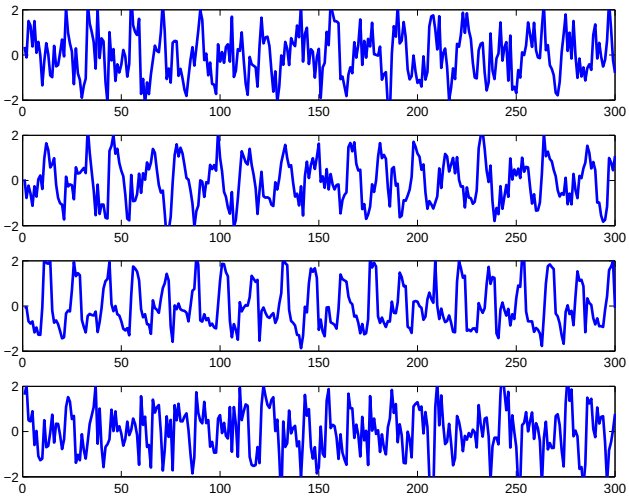

Figura 4.5: Señales sintéticas mezcladas mediante una matriz generada aleatoriamente.

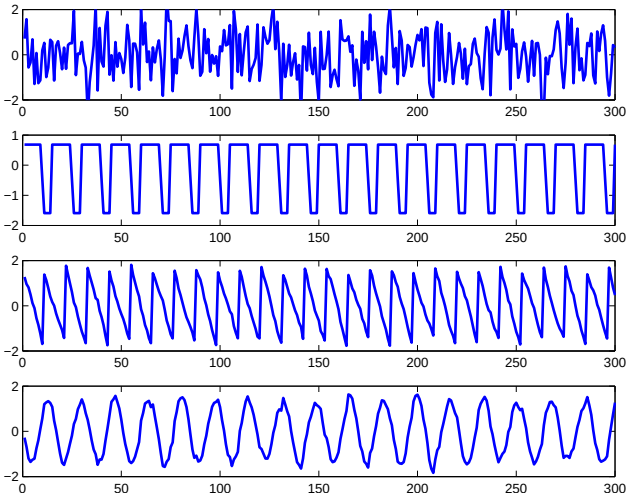

Figura 4.6: Señales separadas por el algoritmo ParzenICA.

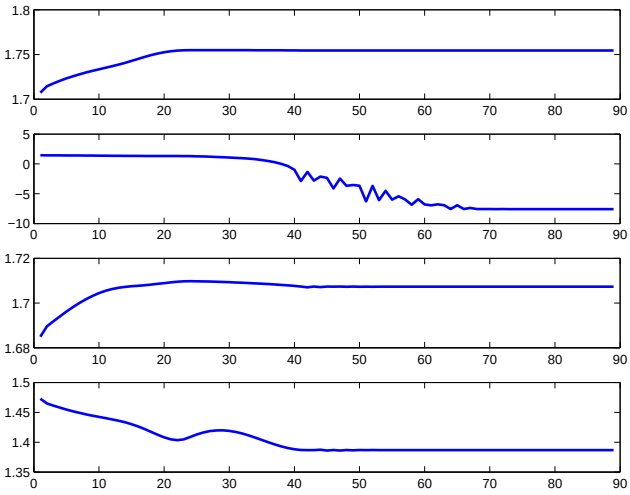

Figura 4.7: Evolución de la entropía de cada proyección.

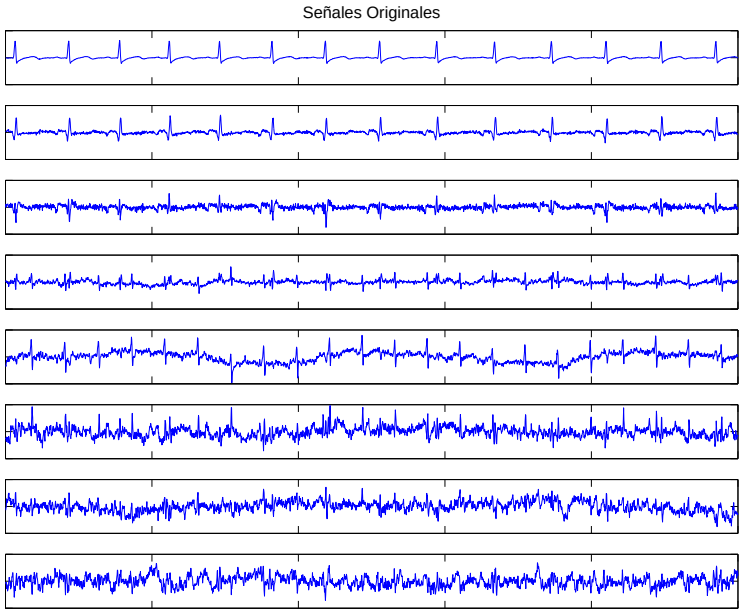

Figura 4.8: Señales ECG originales.

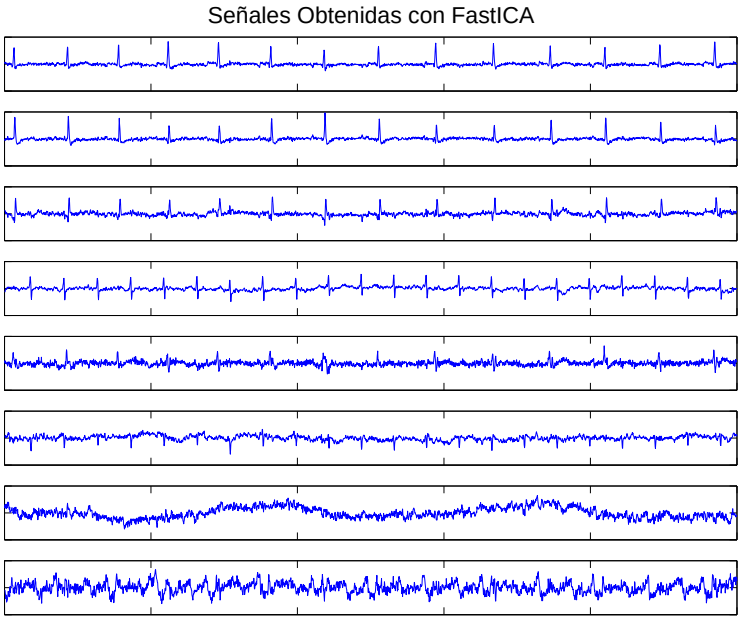

Figura 4.9: Señales ECG separadas por FastICA.

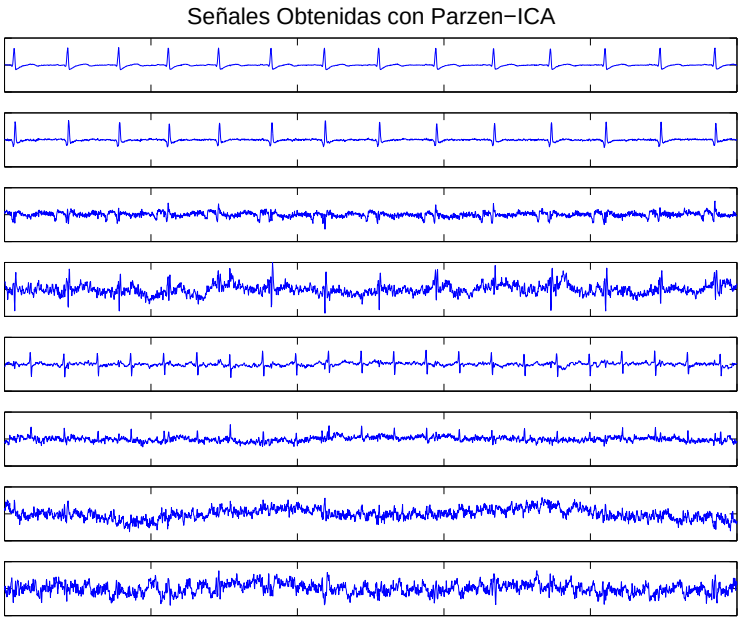

Figura 4.10: Señales ECG separadas por ParzenICA.

de ruido (rizado) ligeramente menor que en el caso de FastICA. En este último caso, las señales correspondientes a los latidos de la madre son las dos primeras y la correspondiente al feto es la cuarta, en la Figura 4.9. Se comprueba, por tanto, la mayor capacidad de ParzenICA respecto a FastICA para separar señales mezcladas linealmente. Además, lo consigue sin el requerimiento de aplicar una decorrelación previa a los datos.

Mediante estas simulaciones, el algoritmo ParzenICA se ha demostrado preciso en su capacidad para llevar a cabo separación ciega de señales, sin el requerimiento de decorrelación previa. Sin embargo, el algoritmo propuesto es de tipo bloque. Esto quiere decir que la complejidad del método depende de la del modelo de Parzen utilizado. Esto hace que ParzenICA no sea viable para separación ciega en línea, lo que requeriría una versión estocástica del método. El problema de la estimación estocástica de la entropía ha sido tratado en [Erdogmus, 2003], y puede constituir una solución al problema de ICA en tiempo real, lo que constituye una propuesta de línea de trabajo futuro como extensión del método aquí presentado.

Capítulo 5

Conclusiones y líneas de trabajo futuras

A lo largo de esta tesis se ha presentado una serie de aportaciones al aprendizaje basado en teoría de información, con especial interés en el problema de la extracción de características. Las dos vías exploradas, una basada en funciones de contraste y otra basada en la estimación de la FDP, se enfrentan a una serie de dificultades a las que se han planteado soluciones parciales.

La primera de estas vías ha sido desarrollada en el Capítulo 2, en la que se ha planteado el método MMI, basado en una función de contraste que trata de aproximar lo mejor posible el valor de la información mutua entre las proyecciones obtenidas y las clases. Esta función de coste ha demostrado ser eficaz para la extracción de la información relevante para clasificación. Aunque en las simulaciones con bases de datos públicas no consigue destacarse significativamente del resto de los métodos de referencia con que se compara, en la prueba con datos sintéticos se demuestra capaz de extraer las características relevantes incluso cuando la función de discriminación entre las clases cuenta con fuertes no linealidades. Por otro lado la estimación de la información mutua propuesta puede servir, además de para la extracción de nuevas características, para evaluar individualmente un conjunto de características en el es-

pacio original. Los resultados presentados en dicho capítulo han sido publicados en [Leiva-Murillo, 2007a]. Por otro lado, la estimación de la información mutua propuesta ha sido aplicada con éxito al problema de la identificación de distintos locutores a partir de una secuencia de audio digital. La segmentación se lleva a cabo mediante una búsqueda global, por medio de un algoritmo de temple simulado, de la secuencia de etiquetas que hace máxima la información mutua entre la secuencia de coeficientes Cepstrum con que se ha codificado el audio y dichas etiquetas [Leiva-Murillo, 2007b]. La aplicación de un algoritmo genético con un propósito similar ha sido publicado en [Salcedo-Sanz, 2006].

El método MMI hace frente a dos fuentes de imprecisión a la hora de estimar la información mutua entre las características extraídas y las clases. Por un lado, el error de la aproximación de las entropías involucradas en la estimación de la información mutua. Por otro lado, el hecho de que estemos considerando la suma de la "informaciones" $\sum_k I(z_k, y)$ en lugar de la información $I(\mathbf{z}, y)$. Las futuras líneas de investigación que traten de incrementar las prestaciones de MMI se han de enfrentar a ambas fuentes de imprecisión. Para ello, se propone profundizar en algunos trabajos presentados en la literatura y que pueden ayudarnos a solventar estas limitaciones. De entre ellas, las que se proponen por su mayor interés son las siguientes:

- Encontrar una estimación robusta de la entropía unidimensional, más precisa que la basada en momentos no polinomiales. En [Welling, 2005] se propone una estimación robusta de la FDP de una variable unidimensional a partir de sus muestras. La estimación de la entropía a partir de ese modelo de FDP no es trivial, por lo que se hace necesario profundizar en el estudio analítico de la obtención de la entropía.

- Encontrar una estimación de la entropía de una variable multidimensional. Se ha propuesto una generalización de la expansión de Edgeworth para la estimación de FDP de variables multidimensionales en [Hulle, 2005]. Se necesita, por

tanto, obtener una estimación de la entropía a partir de ésta.

- Buscar la maximización de la suma $\sum_k I(z_k, y)$ en lugar de $I(\mathbf{z}, y)$ puede ser válido si la diferencia entre las magnitudes permanece bajo control. Lo que se ha propuesto en el planteamiento del método MMI es aplicar PCA para decorrelacionar las variables, y forzar que las proyecciones sean ortogonales y por tanto preservar la decorrelación a la salida. Dado que decorrelación no implica independencia estadística, se deberá buscar la manera de introducir, de forma más fiable, un término de penalización de la información mutua entre las propias variables. La función de coste, en ese caso, debería tener la forma

$$\arg\max_{\mathbf{W}} \sum_k I(z_k, y) - I(\mathbf{z}) + I(\mathbf{z}|y) \tag{5.1}$$

  ya que el término de penalización $I(\mathbf{z}) - I(\mathbf{z}|y)$ consiste en la diferencia entre $\sum_k I(z_k, y)$ e $I(\mathbf{z}, y)$. Se deberá, por tanto, encontrar la forma de estimar el valor de $I(\mathbf{z})$ y de $I(\mathbf{z}|y)$. Para ello, surge una dificultad que impide que puedan usarse técnicas ICA al uso. Por ejemplo, con FastICA se asume que $h(\mathbf{z}) = h(\mathbf{x})$ ya que la transformación es invertible y por tanto se puede despreciar el término. Ahora no es así, porque la transformación no es invertible: la matriz $\mathbf{W}$ ya no es cuadrada como en el caso de ICA.

La segunda línea de trabajo que se ha explorado en esta tesis se ha basado en estimaciones de la FDP conforme a la que están generados los datos. A partir de ella, se estiman magnitudes como la entropía o el test de hipótesis en el espacio transformado. Los resultados obtenidos con clasificadores de Parzen ponen de relieve la fiabilidad de los modelos obtenidos mediante el método ML-LOO, ya que se alcanzan prestaciones competitivas respecto a otros métodos más populares como SVM ó KNN. El clasificador de Parzen no ha tenido un lugar destacado en la literatura sobre reconocimiento de patrones por la dificultad de establecer un ancho adecuado para la ventana que hiciese que los modelos en los que se basa fuesen precisos. El

método ML-LOO aporta un criterio para el diseño del modelo que alivia este problema, además de dar libertad para la elección de una matriz de covarianza para la ventana con distintos grados de complejidad (esférica, diagonal o completa). El Teorema 3.1 aporta la demostración de la convergencia del algoritmo para el caso esférico. El Teorema 3.2 establece, además, la optimalidad del estimador propuesto para el ancho de la ventana en términos de estimación de la entropía cuando se hace mediante un muestreo del tipo *leave-one-out*. Esto es importante para la aplicación de modelos de Parzen a técnicas de aprendizaje basadas en teoría de la información que necesiten una estimación de la entropía.

En cuanto a la extracción de características, el método MaxLT basado en la optimización del test de hipótesis uno-contra-el-resto se ha demostrado igualmente eficaz. Se ha probado su conexión con otro método propuesto en la literatura, el análisis de discriminantes informativos, y se ha mostrado la equivalencia de ambos en cuanto a prestaciones, y la idoneidad de ML-LOO para construir los modelos que conducen a unas aceptables prestaciones sin necesidad de llevar a cabo validación cruzada para determinar el ancho de la ventana de Parzen, lo que reduce considerablemente la complejidad computacional de este tipo de técnicas. A pesar de que el test de hipótesis es planteado como uno-contra-el-resto, los resultados de clasificación cuando la regla de discriminación es del tipo uno-contra-uno son del mismo orden que para uno-contra-el-resto. Junto a este esquema, se ha presentado otro basado en maximización de la información mutua mediante modelos GMM cuyas prestaciones llegan a superar al algoritmo MaxLT. La principal razón por las que el método MMI-GMM supera a MaxLT para reducciones drásticas de la dimensión puede hallarse en el hecho de que la distribución de los datos a la entrada y a la salida del extractor de características está mejor caracterizada en el caso de los GMM, ya que se establece la relación entre los parámetros del modelo para $\mathbf{z}$, dados por $\{\mathbf{m}_k, \mathbf{C}_k\}_{k=1,\cdots,K}$, cosa que no ocurre con el modelo de Parzen usado en MaxLT. Por estas razones, se proponen las siguientes extensiones al trabajo presentado:

- Mejorar la optimización en el algoritmo MaxLT. El hecho de que no alcance

unas prestaciones competitivas respecto a otros métodos del estado del arte puede explicarse por la ineficacia en la correcta elección del ancho de ventana. A medida que evoluciona el algoritmo, este ancho habrá de adaptarse a la estadística de los datos en baja dimensión, que irá variando en el transcurso de la búsqueda de la proyección óptima.

- Proponer una versión de MaxLT que haga uso de modelos de Parzen completos. En este caso, se habrá de buscar la forma de caracterizar la relación entre la matriz de covarianza de la ventana para el espacio de características original y el correspondiente al espacio de dimensión reducida.

- Mejorar el algoritmo MMI-GMM mediante la determinación automática del parámetro K óptimo (número de gausianas o *clusters*) para el modelo GMM de cada clase. Esto puede ser llevado a cabo mediante algún criterio de penalización de la complejidad, como el criterio de Akaike o el criterio de información bayesiano [Burnham, 1998].

- Extender el esquema uno-contra-el-resto, de forma que cada uno de los N_c clasificadores binarios pueda contar con su propia matriz de extracción de características. Es decir, en lugar de una matriz $\mathbf{W}$, buscamos un conjunto de matrices de transformación $\{\mathbf{W}_l\}_{l=1,\cdots,N_c}$, de forma que cada una de ellas sea óptima para el l-ésimo clasificador binario. El diseño conjunto de estas matrices debe dar lugar a un clasificador uno-contra-el-resto que tenga en cuenta las particularidades de cada problema binario en concreto.

En la última de las aplicaciones presentadas en el Capítulo 4, se ha descrito la aplicación del modelo de Parzen optimizado con ML-LOO a separación ciega o ICA, demostrándose la viabilidad del procedimiento descrito para la separación de señales tanto cuando estas son generadas sintéticamente como cuando corresponde a datos reales.

Frente a estas mejoras propuestas para los métodos presentados, se proponen dos líneas de trabajo de mayor profundidad para generalizar los métodos presentados

en esta tesis. Aunque los resultados presentados para los distintos problemas de aprendizaje máquina avalan la validez del estimador de Parzen propuesto, existe una limitación debida al alto coste computacional que supone trabajar con los modelos de Parzen. En ICA, este coste hace impracticable la aplicación del método ParzenICA a problemas en los que la separación deba llevarse a cabo en línea. Para solventar este problema, se propone como primera extensión al trabajo presentado en esta tesis, y que se ha centrado en métodos de tipo bloque, el uso de métodos estocásticos de estimación y manipulación de magnitudes como entropía e información mutua. En segundo lugar, el potencial del método ML-LOO para la precisa estimación de la FDP de los datos puede ser de aplicación para otro de los problemas de aprendizaje más estudiados en la literatura, como es el de agrupamiento o *clustering* mediante modelos de Parzen. A continuación se describen, con mayor detalle, las dos líneas de trabajo mencionadas.

5.1. Métodos estocásticos de aprendizaje

La metodología seguida para los algoritmos presentados en el Capítulo 4 ha sido de tipo bloque: todas las muestras se usan a la vez para la estimación de los cocientes de verosimilitud, y ésta es, a su vez estimada a partir de todas las muestras disponibles. Sería por tanto una ventaja encontrar versiones estocásticas u *on-line* que permitan el entrenamiento de las matrices de extracción de forma recursiva, de modo que una nueva muestra pueda ser incorporada al aprendizaje, otorgando al algoritmo capacidad de adaptatividad a los datos. Se ha propuesto en [Erdogmus, 2003] un método para estimación adaptativa de la entropía que, además de ser aplicable *on-line* y por tanto capaz de captar comportamientos dinámicos en el tiempo por parte de los datos, reduce la complejidad de la estimación. En este caso, la complejidad computacional pasa de $O(N^2)$ a $O(N)$. La aproximación de la entropía

propuesta proporciona un valor de la entropía, para la muestra n-ésima, dado por

$$\hat{h}_n(\mathbf{x}) = -\log \frac{1}{N} \sum_{i=1}^{N} G(\mathbf{x}_n - \mathbf{x}_{n-1}, 2\sigma^2) \tag{5.2}$$

Una interesante línea de trabajo por desarrollar será la de estudiar la aplicación de este estimador a alguno de los problemas que han sido presentados a la largo de la tesis y que han sido resueltos mediante estimadores de tipo bloque.

5.2. Agrupamiento mediante modelos de Parzen

El problema del agrupamiento o *clustering* es, junto con ICA, el más popular de entre los métodos de aprendizaje no supervisado. El problema consiste en encontrar una partición del espacio vectorial en el que residen los datos. También puede interpretarse como una asignación de etiquetas, de entre un conjunto finito de ellas, a un conjunto de muestras. Desde este punto de vista, la única diferencia entre agrupamiento y clasificación es el carácter no supervisado, ya que ahora no contamos con un subconjunto de muestras etiquetadas que nos permita construir una generalización hacia muestras no etiquetadas.

Al igual que el diseño de modelos de Parzen mediante el método ML-LOO puede incrementar notablemente las prestaciones de un clasificador basado en estos modelos, también podría hacerlo con el resultado de un agrupamiento de los datos. Sin embargo, el problema presenta dificultades que no existían en el problema de clasificación. El agrupamiento consiste en encontrar la relación de pertenencia de cada muestra a cada *cluster*. Pero esa relación de pertenencia es también necesaria para el propio diseño de los modelos, ya que el correspondiente a cada grupo, para ser preciso, debe estar construido con las muestras pertenecientes al mismo, para de esta forma determinar el ancho de ventana óptimo.

Desde un punto de vista de teoría de la información, el agrupamiento óptimo será el que asigna al conjunto de datos $\mathbf{X}$ un vector de etiquetas $\mathbf{Y}$ de forma que la información mutua entre ambas sea máxima. Esto lleva a establecer el criterio de

asignación

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}} I(\mathbf{X}, \mathbf{Y})$$

Este criterio de máxima información mutua puede ser transformado en un criterio de mínima entropía de la forma

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}} [h(\mathbf{X}) - h(\mathbf{X}|\mathbf{Y})] = \arg \min_{\mathbf{Y}} h(\mathbf{X}|\mathbf{Y})$$

dado que la entropía $h(\mathbf{X})$ es constante.

El valor de esta entropía viene dado, una vez conocido $\mathbf{Y}$, por

$$\begin{aligned} h(\mathbf{X}|\mathbf{Y}) &= \sum_{l=1}^{L} P(c_l) h(\mathbf{X}|c_l) \\ &= -\sum_{l} \frac{N_l}{N} \sum_{i \in I_l} \frac{1}{N_l} \log\Big(\sum_{j \in I_l} \frac{1}{N_l} G_{ij}\Big) \\ &= -\frac{1}{N} \sum_{l} \sum_{i \in I_l} \log\Big(\sum_{j \in I_l} \frac{1}{N_l} G_{ij}\Big) \end{aligned} \tag{5.3}$$

siendo $G_{ij} = G(\mathbf{x}_i - \mathbf{x}_j|\sigma^2)$. La búsqueda del $\mathbf{Y}$ óptimo es, en principio, un problema NP-completo, ya que exige la evaluación del criterio para todos los vectores $\mathbf{Y}$ posibles. Por tanto, para convertir el problema del agrupamiento en un problema de optimización, necesitamos establecer un esquema de tipo "blando", que asigne a cada muestra un valor de probabilidad de pertenencia a cada uno de los grupos. Si definimos la variable indicadora r_{li} como $r_{li} = p(c_l|\mathbf{x}_i)$, de acuerdo con lo anterior debe cumplirse que $\sum_l r_{li} = 1$. La entropía (5.3) quedaría, en función de estas variables, de la forma

$$h(\mathbf{x}|y) = -\frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N} r_{li} \log\Big(\frac{\sum_{j=1}^{N} r_{lj} G_{ij}}{\sum_{j=1}^{N} r_{lj}}\Big) \tag{5.4}$$

donde puede comprobarse la equivalencia con la expresión (5.3) cuando las r_{li} se saturan, es decir, cuando toman únicamente valores iguales a 1 ó a 0.

El problema de la optimización del funcional (5.4), con las restricciones $r_{li} \geq 0$ y $\sum_l r_{li} = 1$, padece un problema de mínimos locales.

La ventaja de encontrar la forma de minimizar la información mutua entre datos y grupos reside en el hecho de que proporcionaría una flexibilidad de la que carecen los modelos de mezcla de gausianas para hacer frente a datos cuyas muestras se hallan agrupadas mediante estructuras con fuertes no-linealidades. Desde ese punto de vista, el método puede proporcionar buenos resultados tanto para datos agrupables mediante modelos GMM u otras técnicas basadas en distancias euclídeas, como para datos donde predomine la conectividad local en la estructura de los *clusters*, y por tanto se presten más a métodos basados en grafos de conectividad, vecinos más próximos o *clustering* espectral.

Bibliografía

Ahmad, I. y Lin, P. (1976). A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory*, 22(3): 372–375.

Akaho, S. (2002). Conditionally independent component analysis for supervised feature extraction. *Neurocomputing*, 49(1-4): 139–150.

Antos, A., Devroye, L. y Györfi, L. (1999). Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7): 643–645.

Arndt, C. (2001). *Information Measures*. Springer, Berlín.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, 3: 213–251.

Bach, F. y Jordan, M. (2002). Kernel independent component analysis. *Journal on Machine Learning Research*, 3: 1–48.

Backer, S., Naud, A. y Scheunders, P. (1998). Non-linear dimensionality reduction techniques for unsupervised feature extraction. *Pattern Recognition Letters*, 19(8): 711–720.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4): 537–550.

Beirlant, J., Dudewicz, E., Györfy, L. y van der Meulen, E. (1997). Nonparametric entropy estimation. *Int. J. Mathematical and Statistical Sciences*, (6): 17–39.

Bell, A. y Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6): 1004–1034.

Bhattacharyya, C. (2004). Second order cone programming formulations for feature selection. *Journal of Machine Learning Research*, 5: 1417–1433.

Bi, J., Bennett, K., Embrechts, M., Breneman, C. y Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3: 1229–1243.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Blum, A. y Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2): 245–271.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71: 353–360.

Burnham, K. y Anderson, D. (1998). *Model Selection and Inference: a Practical Information-Theoretic Approach*. Springer, New York.

Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation*, (10): 2009–2025.

Chow, T. y Huang, D. (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *IEEE Transactions on Neural Networks*, 16(1): 213–224.

Cover, T. y Thomas, J. (2006). *Elements of Information Theory, 2nd Edition*. John Wiley & Sons, Hoboken, NJ.

Duda, R. y Hart, P. (2000). *Pattern Classification*. John Wiley & Sons, New York.

Duin, R. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, 25(11).

Erdogmus, D., Hild, K. y Principe, J. (2003). Online entropy manipulation: Stochastic information gradient. *IEEE Signal Processing Letters*, 10(8): 242–245.

Erdogmus, D. y Príncipe, J. (2004). Lower and upper bounds for misclassification probability based on Renyi's information. *Journal of VLSI Signal Processing Systems archive*, 37(2–3): 307–317.

Feng, D., Zheng, W. y Jia, Y. (2005). Neural network learning algorithms for tracking minor subspace in high-dimensional data stream. *IEEE Transactions on Neural Networks*, 16(3): 513–521.

Fletcher, R. (1995). *Practical Methods of Optimization (2nd Edition)*. John Wiley & Sons, New York.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5: 1531–1555.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Fukunaga, K. y Hayes, R. (1989). The reduced Parzen classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4): 423–425.

Goldberger, J., Roweis, S., Hinton, G. y Salakhutdinov, R. (2005). Neighbourhood components analysis. En *Advances in Neural Information Processing Systems, NIPS*, pp. 513–520, Vancouver, Canadá.

Guyon, I. y Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3: 1157–1182.

Guyon, I., Weston, J., Barnhill, S. y Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46: 389–422.

Hall, P. (1982). Cross-validation in density estimation. *Biometrika*, 69(2): 383–390.

Hardoon, D., Szedmák, S. y Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12): 2639–2664.

Hastie, T. Tibshirani, R. y Friedman, J. (1998). *The Elements of Statistical Learning*. Springer, New York.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, Upper Saddle River.

Heisele, B., Poggio, T. y Pontil, M. (2000). Face detection in still gray images. Technical Report 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA.

Hellman, M. y Raviv, J. (1970). Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, 16(4): 368–372.

Hild, K., Erdogmus, D., Torkkola, K. y Príncipe, J. (2006). Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9): 1385–1392.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28: 321–377.

Hsieh, P. y Landgrebe, D. (1998). Linear feature extraction for multiclass problems. En *IEEE Int. Geoscience and Remote Sensing Symp.*, volumen 4, pp. 2050–2052, Seattle, USA.

Hsieh, P., Wang, D. y Hsu, C. (2006). A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant infor-

mation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 28(2): 223–235.

Huber, P. (1985). Projection pursuit. *Annals of Statistics.*, 13(2): 435–475.

Hulle, M. M. V. (2005). Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9): 1903–1910.

Hyvarinen, A. (1998). New approximations of differential entropy for independent component analysis. volumen 10, pp. 273–279, Vancouver.

Hyvarinen, A., Karhunen, J. y Oja, E. (2001). *Independent Component Analysis.* John Wiley & Sons, New York.

Jeon, B. y Landgrebe, D. (1994). Fast Parzen density estimation using clustering-based branch and bound. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9): 950–954.

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute for Statististical Mathematics*, 41(4): 683–697.

Jones, M. C., Marron, J. S. y Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433): 401–407.

Kamimura, R. (2006). Cooperative information maximization with Gaussian activation functions for self-organizing maps. *IEEE Transactions on Neural Networks*, 17(4): 909–918.

Kaski, S. y Peltonen, J. (2003). Informative discriminant analysis. En *International Conference on Machine Learning, ICML*, pp. 329–336. AAAI Press.

Kay, S. (1998). *Fundamentals of Statistical Signal Processing. Volumen II, Detection Theory.* Prentice-Hall, New York.

Kobayashi, K. y Komaki, F. (2006). Information criteria for support vector machines. *IEEE Transactions on Neural Networks*, 17(3): 571–577.

Kohavi, R. y John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97: 273–324.

Kohonen, T. (2001). *Self-Organizing Maps, 3rd Edition*. Springer, Berlín.

Koller, D. y Sahami, M. (1996). Toward optimal feature selection. En *International Conference on Machine Learning, ICML*, pp. 284–292, Bari, Italia.

Kraskov, A., Stoegbauer, H. y Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69: 066138.

Kwak, N. y Choi, C. (2002a). Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12): 1667–1671.

Kwak, N. y Choi, C. (2002b). A new method of feature extraction and its stability. En *International Conference on Artificial Neural Networks, ICANN*, volumen 2415, pp. 480–485, Madrid, Spain.

Land, A. y Doig, A. (1960). An automatic method of solving discrete programming problems. *Econometrica*, 28: 497–520.

Lee, D. y Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788–791.

Leiva-Murillo, J. y Artés-Rodríguez, A. (2006). A fixed-point algorithm for finding the optimal covariance matrix in kernel density modeling. En *International Conference on Acoustic, Speech and Signal Processing*, pp. V–705 – V–708, Toulouse, Francia.

Leiva-Murillo, J. y Artés-Rodríguez, A. (2007a). Maximization of mutual information for supervised linear feature extraction. *IEEE Transactions on Neural Networks*, 18(5): 1433–1441.

Leiva-Murillo, J., De-Prado, M. y Artés-Rodríguez, A. (2004). Feature extraction for SVMs. En *Learning 2004*, Elche, España.

Leiva-Murillo, J., Salcedo-Sanz, S., Gallardo-Antolín, A. y Artés-Rodríguez, A. (2007b). A simulated annealing approach to speaker segmentation in audio databases. *Engineering Applications of Artificial Intelligence*. Aceptado para su publicación.

Leray, P. y Gallinari, P. (1999). Feature selection with neural networks. *Behaviormetrika*, 26.

Li, H., Jiang, T. y Zhang, K. (2006). Efficient and robust feature extraction by maximum margin criterion. *IEEE Transactions on Neural Networks*, 17(1): 157–165.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Society*, 86(414): 316–327.

Li, K. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420): 1025–1039.

Linsker, R. (1988). Self–organization in a perceptual network. *Computer*, pp. 105–117.

Loog, M. y Duin, R. (2004). Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6): 732–739.

Loog, M., Duin, R. y Haeb-Umbach, R. (2001). Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7): 762–766.

Mao, K. (2003). Fast orthogonal forward selection algorithm for feature subset selection. *IEEE Transactions on Neural Networks*, 13(5): 1218–1224.

Martínez, A. y Zhu, M. (2005). Where are linear feature extraction methods applicable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12): 1934–1944.

Moor, B., De-Gersem, P., De-Schutter, B. y Favoreel, W. (1997). DAISY: Database for the identification of systems. Technical report, Univ. Católica de Lovaina, Dpt. Signals, Identification, System Theory and Automation.

Newman, D., Hettich, S., Blake, C. y Merz, C. (1998). UCI repository of machine learning databases. Informe técnico, Univ. of California, Dept. ICS.

Painter, T. y Spanias, A. (2000). Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4): 451 – 515.

Peltonen, J. y Kaski, S. (2005). Discriminative components of data. *IEEE Transactions on Neural Networks*, 16(1): 68–83.

Peng, H., Long, F. y Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Neural Networks*, 27(8): 1226–1238.

Plumbley, M. D. (2004). Lie group methods for optimization with orthogonality constraints. En *5th International Conference, ICA 2004, Granada, Spain*, volumen 3195 of *Lecture Notes in Computer Science*, pp. 1245–1252. Springer.

Principe, J., Xu, D. y Fischer, J. (2000). *Information-Theoretic Learning*, volumen 1 of *Unsupervised Adaptive Filtering*. John Wiley & Sons, New York.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3: 1357–1370.

Rose, K. (1999). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11): 2210–2239.

Roweis, S. y Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290: 2323–2326.

Salcedo-Sanz, S., Gallardo-Antolín, A., Leiva-Murillo, J. y Bousoño-Calzón, C. (2006). Offline speaker segmentation using genetic algorithms and mutual information. *IEEE Transactions on Evolutionary Computation*, 10(2): 175–186.

Scholkopf, B. y Smola, A. (2002). *Learning with Kernels*. The MIT Press, Cambridge, MA.

Schölkopf, B., Smola, A. y Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(15): 1299–1319.

Scott, D. (1992). *Multivariate Density Estimation*. John Wiley & Sons, New York.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, Londres.

Tenenbaum, J., de Silva, V. y Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319–2323.

Tipping, M. E. y Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3): 611–622.

Tishby, N., Pereira, F. y Bialek, W. (1999). The information bottleneck method. En *37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, Urbana, Illinois.

Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *Journal on Machine Learning Research*, 3: 1415–1438.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory, Part I.* John Wiley & Sons, New York.

Vapnik, V. (1998). *Statistical Learning Theory.* John Wiley & Sons, New York.

Welling, M. (2005). Robust higher order statistics. En *International Conference on Artificial Intelligence and Statistics, AISTATS 2005*, pp. 405–412, Barbados.

Weston, J., Elisseeff, A. y Schölkopf, B. (2001a). Use of the zero-norm with linear models and kernel methods. *Journal on Machine Learning Research*, 3: 1439–1461.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. y Vapnik, V. (2001b). Feature selection for SVMs. En *Advances in Neural Information Processing Systems, NIPS*, pp. 668–674, Vancouver, Canadá.

Yang, Z. y Zwolinski, M. (2001). Mutual information theory for adaptive mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4): 396–403.

Yu, L. y Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5: 1205–1224.